

ANOMALIES IN LINK MINING BASED ON MUTUAL INFORMATION

ZAKEA IDRIS ALI IL-AGURE

A thesis submitted in partial fulfilment of the requirement of Staffordshire University for the
degree of Doctor of Philosophy

July 2015

Acknowledgments

Foremost, I would like to express my sincere gratitude to my principal supervisor Prof. Bernadette Sharp for the continuous support of my PhD study and research, for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my PhD study.

Besides my supervisor, I would like to thank the rest of my thesis committee, Dr. Clare Stanier for her encouragement, insightful comments, and hard questions.

Many thanks go to Dr. Emily Raeburn who has provided me with invaluable advice on statistical analysis.

A special thanks to my brother Dr. Ali Idris. Words cannot express how grateful I am to have you as a brother, for helping me survive all the stress from my PhD journey and not letting me give up.

Last but not the least, I would like to thank my family, my husband Shakier Khalifa for all the sacrifices that he has made on my behalf, for his encouragement and for pushing me further than I thought I could go. I would like to thank my daughters, Shahed and Nour who supported me to strive towards my goal. Finally my son, Shehab, who has sustained me thus far and made this a very fun and pleasurable journey.

Abstract

The literature review found surprisingly low utilisation of mutual information in detecting anomalies in various domains, however no such study in link mining was found. This research is intended to fill the gap in link mining domain, although it has been widely used in other areas of data analysis. The current study is a first-step exploration of a new method that uses mutual information based measures to interpret anomalies and link strength between individual anomalies in a given dataset. Anomalies detection, which is the focus of this research proposal, is concerned with the problem of finding non-conforming patterns in datasets. This thesis describes a novel approach to measure the amount of information shared between any random anomaly variables. Two types of data were used to evaluate the proposed approach: proof of concept data in Case study 1 and citation data in Case study 2. The CRISP data mining methodology was updated to be applicable for link mining study. The proposed mutual information approach to provide a semantic investigation of the anomalies and the updated methodology can be used in other link mining studies such as fraud detection, network intrusion detection and law enforcement areas which are expected to grow.

Keywords: Link mining, anomalies, mutual information and co-citation data.

Table of contents

Abstract	ii
List of Tables	3
List of Figures	3
1 Introduction.....	4
1.1 Key issues of this research.....	5
1.2 Aim and objectives	7
1.3 Research methodology	9
1.4 Ethics.....	10
1.5 Research contributions	10
1.6 Thesis structure.....	10
2 Link Mining and Anomalies Detection	12
2.1 Emergence of link mining	12
2.2 Link mining tasks	13
2.2.1 Object-related tasks.....	13
2.2.2 Graph-related tasks	13
2.2.3 Link-related tasks	13
2.3 Link mining challenges.....	15
2.4 Applications of link mining.....	16
2.5 Anomalies detection	17
2.6 Anomalies detection approaches and methods	18
2.6.1 Nearest neighbour based detection techniques.....	20
2.6.2 Clustering-based anomalies detection techniques	21
2.6.4 Classification techniques	24
2.6.5 Information Theory Based.....	25
2.6.6 Other Techniques.....	25
2.6.7 Overview of strengths and limitations	26
2.7 Challenges of anomalies detection	27
2.8 Anomalies detection and link mining	28
2.9 Summary	28
3 Anomalies in link mining based on mutual information	29
3.1 Mutual Information in Information Theory.....	29
3.1.1 Estimation of mutual information	30
3.1.2 Entropy vs. mutual information	31
3.1.3 Applications of mutual information	32
3.2 Proposed novel approach	35
3.3 Methodology of link mining.....	35
3.3.1 Knowledge Discovery Databases (KDD)	36
3.3.2 SEMMA.....	37
3.3.3 CRISP-DM	38
3.4 Link Mining Methodology	40
3.5 Summary	42
4 Anomalies Detection: Case study 1	44
4.1 Overview of Case study1	44
4.2 Anomaly detection methodology applied to Case study 1	44
4.2.1 Stage 1: Data description	45
4.2.2 Stage 2: Data pre-processing.....	46

4.2.3 Stage 3: Data transformation/coding.....	47
4.2.4 Stage 4: Data exploration	47
4.2.5 Stage 5: Data modelling.....	51
4.2.6 Stage 6: Data evaluation.....	58
4.3 Discussion	59
4.4 Summary	61
5 Anomalies Detection: Case Study 2.....	63
5.1 Anomaly detection methodology applied to Case study 2	63
5.1.1 Stage1: Data description	64
5.1.2 Stage 2: Data pre-processing.....	64
5.1.3 Stage 3: Data transformation	65
5.1.4 Stage 4: Data exploration	69
5.1.5 Stage 5: Data modelling.....	71
5.1.5.1 Graph analysis of co-citation data.....	73
5.1.5.2 Hierarchical Cluster	74
5.1.5.3 Visualisation.....	77
5.1.6 Stage 6: Data evaluation.....	78
5.2 Discussion	80
5.3 Summary	82
6 Conclusion and future work.....	84
6.1 Introduction	84
6.2 Evaluations of the main approach	84
6.2.1 Finding of Case study1	86
6.2.2 Finding of Case study 2	87
6.3 Research contributions	87
6.3.1 Major contributions	87
6.3.2 Minor contributions.....	89
6.4 Limitations of the study	89
6.5 Challenges	89
6.6 Future work	90
7 References	92
Glossary	106
Appendix A	109
Appendix B	112
Appendix C	118
Appendix D	132
Appendix E.....	145
Appendix F.....	147

List of Tables

<i>Table 3.1. Summary of differences between KDD, CRISP-DM and SEMMA</i>	<i>40</i>
<i>Table 4. 1 A small sample of the case 1 data</i>	<i>45</i>
<i>Table 4. 2 Case processing summary</i>	<i>46</i>
<i>Table 4. 3 Shows descriptive statics</i>	<i>48</i>
<i>Table 4. 4 Shows the frequency statics</i>	<i>48</i>
<i>Table 4. 5 Shows the frequency of anomalies statistics</i>	<i>48</i>
<i>Table 4. 6 Shows the frequency of non-anomalies/anomalies in the purchase category</i>	<i>49</i>
<i>Table 4. 7 shows the frequency of anomalies/anomalies in staff ID.....</i>	<i>50</i>
<i>Table 4. 8 Shows the number of non anomalies/ anomalies in staff trainer</i>	<i>50</i>
<i>Table 4. 9 Model summary.....</i>	<i>52</i>
<i>Table 4. 10 The cluster view</i>	<i>54</i>
<i>Table 4. 11 List of anomalies in the proof of concept data... ..</i>	<i>56</i>
<i>Table 4.12 Measure of mutual information between two variables.....</i>	<i>61</i>
<i>Table 5. 1 Table of link strength</i>	<i>77</i>
<i>Table 5. 2 Result of mutual information.....</i>	<i>79</i>

List of Figures

<i>Figure 1. 1 Research process steps</i>	<i>7</i>
<i>Figure 1. 2 Research methodology</i>	<i>9</i>
<i>Figure 2. 1 Link mining tasks and challenges</i>	<i>15</i>
<i>Figure 2. 2 methods of anomalies detection.....</i>	<i>19</i>
<i>Figure 3. 1 Venn diagram showing entropy, conditional entropy and mutual information</i>	<i>32</i>
<i>Figure 3. 2 CRISP-Data mining methodology (Shearer 2000)</i>	<i>36</i>
<i>Figure 3. 3 Link mining methodology</i>	<i>41</i>
<i>Figure 4. 1 Link mining methodology.....</i>	<i>44</i>
<i>Figure 4. 2 Anomalies in proof of concept data.....</i>	<i>46</i>
<i>Figure 4. 3 Bar chart of transaction values.....</i>	<i>49</i>
<i>Figure 4. 4 Staff trainer data</i>	<i>51</i>
<i>Figure 4. 5 Cluster sizes</i>	<i>52</i>
<i>Figure 4. 6 Important features</i>	<i>53</i>
<i>Figure 4. 7 Cluster comparison.....</i>	<i>55</i>
<i>Figure 4. 8 Comparison features in cluster 4</i>	<i>57</i>
<i>Figure 4. 9 Mutual information of transaction values, staff trainer and staff ID.....</i>	<i>59</i>
<i>Figure 5. 1 Link mining methodology.....</i>	<i>57</i>
<i>Figure 5. 2 Convert to dialog format</i>	<i>60</i>
<i>Figure 5. 3 Extracting data from CD-fields (citation-documents)</i>	<i>61</i>
<i>Figure 5. 4 Retaining first authors' initials</i>	<i>68</i>
<i>Figure 5. 5 Convert upper/lower cases.....</i>	<i>68</i>
<i>Figure 5. 6 The frequency</i>	<i>69</i>
<i>Figure 5. 7 Making co-citations</i>	<i>70</i>
<i>Figure 5. 8 Making co-citations pairs.....</i>	<i>70</i>
<i>Figure 5. 9 Mapping nodes.....</i>	<i>66</i>
<i>Figure 5. 10 Cases of node anomaly</i>	<i>68</i>
<i>Figure 5. 11 Clustering.....</i>	<i>76</i>
<i>Figure 5.12 Validating the approach</i>	<i>70</i>

1 Introduction

Link mining is a new emerging research area, which differs from data mining. Whilst data mining aims at discovering new potentially hidden patterns in datasets, link mining considers datasets as a linked collection of interrelated objects and therefore it focuses on discovering explicit links between objects. A crucial step in both data and link mining is to ensure that the analysis is undertaken on reliable, robust and efficient data, and to identify outliers, which are observations that are numerically distant from the rest of the data. Reliability of detection anomaly should achieve high data delivery reliability unless the quality of the underlying links makes that infeasible. Robustness should be robust against huge or complex social networks failures, dynamic networks, and topology changes. In spite of these dynamics, it should function without much tuning or configuration. Efficiency in communication often applies both complex anomalies and different types of anomalies, to allow an opportunity to make the method detection anomalies more efficient. Though outliers are often considered as an error or noise in data mining, they are often referred to as anomalies in link mining as they can carry important information. Often the data contains noise that tends to be similar to the actual anomalies and hence it is difficult to distinguish and remove them (Chandola *et al.*, 2009). Any errors in data are to be examined taking into consideration the context of the domains; some may be true errors and therefore removed, whereas other errors may be regarded as interesting anomalies.

Link mining applications have been shown to be highly effective in addressing many important business issues such as money laundering (Kirkland *et al.*, 1999), telephone fraud detection (Fawcett and Provost 1999), crime detection (Sparrow 1991), terrorism (Badia and Kantardzic 2005, Skillicorn 2004), the financial domain (Creamer and Stolfo 2009), social networks and health care problems (Provana *et al.*, 2010, Wadhah *et al.*, 2011). The identification of anomalies is affected by various factors, many of which are of interest for practical applications. For example, criminal deception or fraud will constantly be a costly issue for many profit organisations. Link mining can minimise some of these losses by making use of the massive collections of customer data (Phua *et al.*, 2004) Using web log files, it becomes possible to recognise fraudulent behaviour, changes in behaviour of customers, or faults in systems. Anomalies arise by reasons of such incidents. Consequently, typical fault detection can discover exceptions in the type of items purchased, the amount of

money spent, the time and the location of this purchase information such as the name of the credit holder account number and expiry date which are very easy to obtain, even from one's home mailbox or from any online transaction carried out (Alfuraih *et al.*, 2002). Such automatic systems aimed at preventing fraudulent use of credit cards; detecting unusual transactions are therefore desirable.

Knowledge discovery is the non-trivial removal of implicit, previously unknown, and potentially useful information from data. The type of knowledge that is discovered from databases and its corresponding representational form varies widely depending on both the application area and the database type, such as *data mining*, *text mining*, *web mining* and *link mining*. The specification of the type of knowledge to be discovered directs the pattern-filtering process. *Data mining* involves the use of complicated data analysis tools to discover previously unknown, relationships and valid patterns in large data sets. These tools involve mathematical algorithms, machine-learning methods and statistical models, and applications such as banking, insurance and medicine; while *text mining* has been applied to semi-structured and unstructured information, such as digital libraries and biological information systems. Technologies in the text-mining process include information extraction, topic tracking, summarisation, categorisation, clustering, and concept linkage information extraction (Chakrabarti, 2001). *Web mining* is the extraction of interesting and potentially useful patterns and implicit information from activity related to the World Wide Web whereas *link mining*, focuses on discovering explicit links between objects.

Anomalies detection, which is the focus of this research proposal, is concerned with the problem of finding non-conforming patterns in data sets, such as social network, bibliometrics data and citation. Anomalies can include exceptions, outliers, aberrations, surprises, peculiarities, and so on (Chandola *et al.*, 2009). In data, text and link mining, the first task is to pre-process the data to explore their integrity. Any errors observed in the data, must be analysed within the context of domains and purpose of the analysis.

1.1 Key issues of this research

The problem of detecting anomalies has been studied in particular from a statistical perspective. The user had to model data points using statistical distribution, and points were determined to be anomalies depending on how they appeared in relation to the model. The main problem with these methods is that, in different situations, the user could simply not

have enough knowledge about the underlying data distribution (Ramaswamy *et al.*, 2000). Anomalies can be removed or considered separately in regression modelling to improve accuracy, which can be considered a benefit of anomalies. Identifying them prior to modelling and analysis is important (Williams *et al.*, 2002).

Regression modelling consists in finding a dependence of one random variable or group of variables on another variable or group of variables. In the context of the anomalies-based association method, anomalies are observations that are clearly different from any other points. Once a collection of points has common characteristics, and these common characteristics are ‘anomalies’, these points are associated (Lin & Brown, 2003). Another topic related to anomalies detection is novelty detection (Markou & Singh, 2003a, 2003b; Saunders & Gero, 2000), which aims at detecting previously unobserved (emergent, novel) patterns in the data. The difference between novel patterns and anomalies is that novel patterns are typically incorporated into the normal model after being detected. Many data-mining algorithms find anomalies as a side-product of clustering algorithms; hence, clustering aims to partition a set of *data objects* into a predefined number of clusters. Objects with similar features should be grouped together and objects with different features placed in divided groups (Fränti & Kivijärvi, 2000). However, these techniques define anomalies as points that do not link in clusters. Thus, the proposed novel technique implicitly defines anomalies as the setting noise in which the clusters are embedded, taking into consideration the context of the problem domain.

Another class of techniques defines anomalies as points that are neither part of a cluster nor part of the setting noise; rather, they are special points that behave very differently from the standard (Aggarwal & Yu, 2001). The approach is to choose the clustering that shares the most information with all the other clusterings, (Strehl and Ghosh 2002). A measure is therefore needed to quantify the amount of information shared between clusterings; hence, information theoretic measures from another fundamental class. Such measures work because of their strong mathematical foundation, and their ability to detect non-linear similarities. This class of measures has been popularised through the works of Strehl and Ghosh (2002) and Meila (2005), and has featured in various subsequent research projects (Fern & Brodley, 2003; He *et al.*, 2008; Tumer and Agogino, 2008).

The proposed novel anomaly detection method advocates the use of mutual information to study the relationships between clusters in order to identify vital hidden information in link

mining applications. The novel method is applied to two new areas: transaction data, and citation data.

1.2 Aim and objectives

The aim of the research is to develop a novel approach to provide a semantic interpretation of anomalies based on mutual information.

To achieve the above aims the following objectives are identified, shown in Figure 1.1

1. Formulating the research context.
2. Conducting a literature review related to link mining and anomalies detection.
3. Developing the conceptual method for investigating the use of mutual information to interpret anomalies.
4. Undertaking an exploratory Case study 1.
5. Applying and validating the results of the novel approach on Case study 2.
6. Evaluating the research project.
7. Writing the thesis and publishing findings.

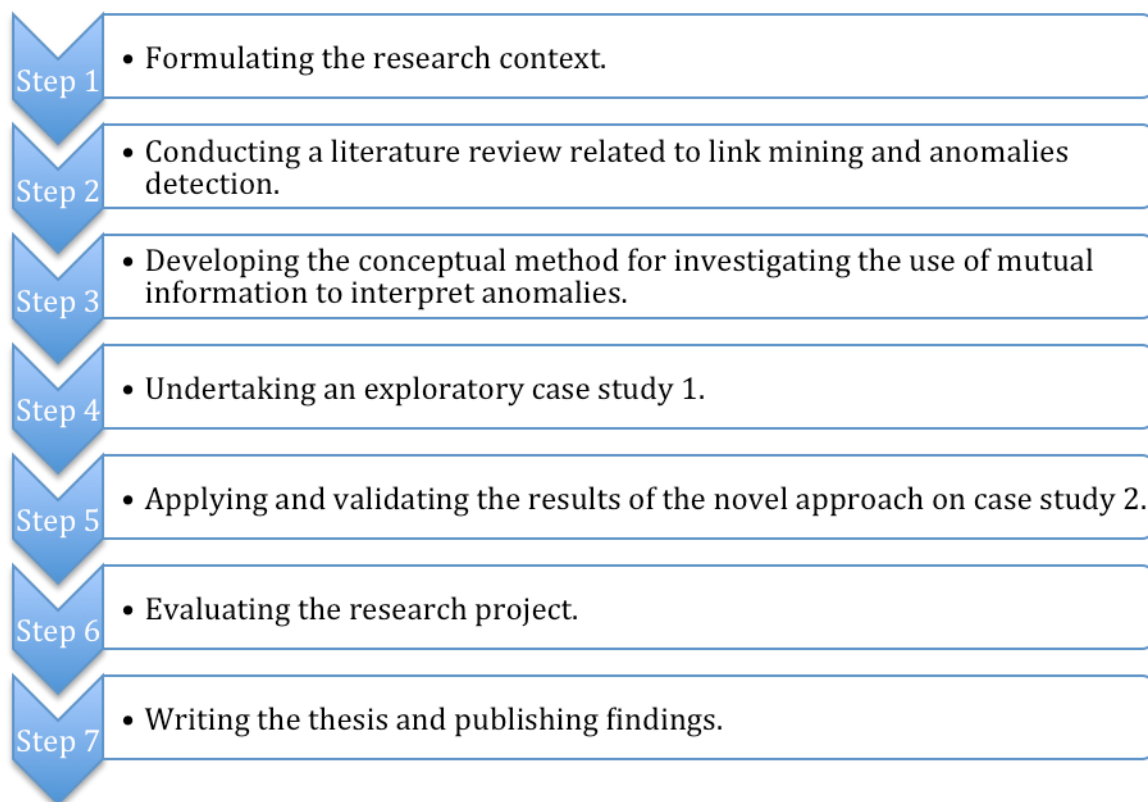


Figure 1. 1 Research process steps

Step 1. Formulating the research context.

This research aims to develop a novel method for detecting anomalies in data sets related to link mining. The anomalies take into consideration the context of the data sets and apply mutual information to measure what object/data item X shares with another object/data item Y .

Step 2. Conducting a literature review related to link mining and anomalies detection.

A literature survey of the current research issues and techniques in link mining is to be carried out. Applications of anomalies detection are to be analysed in order to survey appropriate methods. To investigate the links between objects and understand the context of their anomalies.

Step 3. Developing the conceptual method for conducting the research.

This step focuses on the investigation of the mutual information in link mining and its application to anomalies detection. And adapted CRISP-DM to link mining.

Step 4. Undertaking an exploratory Case study 1.

The first Case study is used as proof of concept to examine the validity of the proposed approach. The mutual information is applied to case 1 to understand/explain the anomalies approach using two -step clustering.

Step 5. Applying and validating the results of the novel approach to Case study 2.

The second Case study uses a set of co-citation data extracted from three databases: SCI-EXPANDED, SSCI, A&HCI. It used BibExcel to analyse the citation data and create a subset of co-citations, using a graph consisting nodes and edges, and use hierarchical clustering provided by VOSviewer to visualise the data. Mutual information is applied to validate the visualisation and to provide a semantic understanding of the anomalies.

Step 6. Overall evaluation of the research project.

This step analyses the validity of the research methodology and the applications of mutual information to the semantic interpretation of anomalies in the data.

Step 7. Writing the thesis and publishing findings.

Seminars were presented to the research community at Staffordshire University and Glyndŵr University. A paper has been submitted to IEEE technically Co-Sponsored SAI intelligent system conference 2015.

1.3 Research methodology

There are three common research approaches: qualitative, quantitative and mixed methods. Qualitative research is described as an unfolding model that occurs in a natural setting that enables the researcher to develop a level of detail from high involvement in the actual experiences (Creswell, 1994) whereas quantitative research begins with a problem statement and involves the formation of a suggestion, a quantitative data analysis and a literature review (Creswell, 2003). ‘A quantitative research relates meaning through objectivity uncovered in the collected data’ (Creswell, 2003, p.19). The mixed methods approaches are an addition of, rather than a replacement, for the qualitative and quantitative approaches to research, as both quantitative and qualitative research continues to be useful and important (Johnson and Onwuegbuzie, 2004).



Figure 1. 2 Research methodology

A quantitative research methodology is used in this research (see Figure 1. 2). Quantitative research is an objective, formal, systematic procedure in which numerical data are used to obtain information. In this method; it is used to describe variables, examine relationships among them and determine the cause-and-effect interactions between these variables (Burns and Grove 2005:23).

This research methodology reviews tasks and challenges as well as current techniques related to link mining. In the first phase of the investigation, the *feature selection* will focus on selecting relevant features for analysis using clustering methods. The next stage determines the best *clustering algorithm* type (i.e. hierarchical, exemplar or conceptual clustering). It

applies mutual information to interpret anomalies found in the data set. It will be used to quantitatively analyse the relationship between any two features, or between a feature and a class variable.

1.4 Ethics

This research presents very limited ethical issues as it does not involve human or animal participants, and does not re-use previously collected personal data. The data used in Case study 1 is proof of concept data designed to test the approach and is fictitious. The data used in Case study 2 is freely available data, in the public domain. Case study 2 is a set of co-citation extracted from three databases: SCI-EXPANDED, SSCI, A&HCI. This research complies with the regulations of Staffordshire University.

1.5 Research contributions

There are three main novel contributions. Link mining is a new emerging research area with applications related to predicting or describing links and relationships among data instances in order to discover patterns in data. This research extends the original purpose of the link mining and attempts at investigating and detecting anomalies in links and relationships among data objects. This attempt extends the initial tasks described by Getoor (2005). This thesis extends the use of link mining by applying mutual information to interpret anomalies. To our knowledge, there is no formal methodology developed in link mining. This research has extended the common CRISP methodology used in data mining and adapted it to link mining.

1.6 Thesis structure

This thesis is organised as follows. The first chapter explores the key problem issues, aim and objectives of the research, its research methodology, ethics and novel contributions. The second chapter focuses more on the concepts and methods of link mining, and reviews anomalies detection techniques and approaches. The third chapter presents the basic concepts of mutual information and addresses the application of mutual information in link mining to detect anomalies. It also describes how the methodology of CRISP-Data Mining can be adapted to link mining. The fourth chapter investigates the use of mutual information to the detection of anomalies using a case study 1 as a proof of concept data. The fifth chapter

applies mutual information citation data Case study 2. The final chapter reviews the proposed novel approach and identifies limitations of the study and proposes future work.

2 Link Mining and Anomalies Detection

This chapter introduces the emergence of link mining and its relevant application to detect anomalies which can include events that are unusual, out of the ordinary or rare, unexpected behaviour, or outliers.

2.1 Emergence of link mining

Link mining is a newly developed research area, bringing together research insights from the fields of web mining, graph theory and machine learning. Link mining applications have been shown to be highly effective in addressing many important business issues such as telephone fraud detection (Fawcett & Provost, 1999), crime detection (Sparrow, 1991), money laundering (Kirkland *et al.*, 1999), terrorism (Badia & Kantardzic, 2005; Skillicorn, 2004), financial applications (Creamer & Stolfo, 2009), social networks and health care problems (Provana *et al.*, 2010; Wadhah *et al.*, 2011). The trend in the building and use of link mining models for critical business, law enforcement and scientific decision support applications are expected to grow. An important issue will be building models and techniques that are scalable and reliable.

Link mining attempts to build predictive or descriptive models of the linked data (Getoor & Diehl, 2005). The term ‘link’ in the database community differs from that in the AI community. In this research a link refers to some real-world connection between two entities (Senator, 2005). Link mining focuses on techniques that explicitly consider these links when building predictive or descriptive models of the data sets (Getoor, 2005). In data mining, the main challenge is to tackle the problem of mining richly structured heterogeneous data sets. The data domains often consist of a variety of object types; these objects can be linked in a variety of ways. Traditional statistical inference procedures assume that instances are independent and this can lead to unsuitable conclusions about the data. However, in link mining, object linkage is a knowledge that should be exploited. In many applications, the facts to be analysed are dynamic, so it is important to develop incremental link mining algorithms, besides mining knowledge from link objects and networks (Getoor & Diehl, 2005).

2.2 Link mining tasks

In their paper, Getoor and Diehl (2005) identify a set of link mining tasks (see Figure 2.1), which are:

- § Object-related tasks.
- § Graph-related tasks.
- § Link-related tasks.

2.2.1 Object-related tasks

These tasks include link-based object clustering, link-based object classification, object identification and object ranking. In a bibliographic domain, the objects include papers, authors, institutions, journals and conferences. Links include the paper citations, authorship and co-authorship, affiliations, and the relation between a paper and a journal or conference.

2.2.2 Graph-related tasks

These tasks consist of sub-graph discovery, graph classification, and generative models for graphs. The aim is to cluster the nodes in the graph into groups sharing common characteristics. In the bibliographic domain, an example of graph classification is predicting the category of a paper, from its citations, the papers that cite it, and co-citations (papers that are cited with this paper).

2.2.3 Link-related tasks

These tasks aim at predicting the existence of a link between two entities based on the attributes of the objects and other observed links. In a bibliographic domain, predicting the number of citations of a paper is an indication of the impact of a paper— papers with more citations are more likely to be seminal.

Link prediction is defined as inferring the existence of a link (relationship) in the graph that is not previously known. Examples include predicting links among actors in social networks, such as predicting friendships or predicting the participation of actors in events (O'Madadhain et al., 2005) such as email, telephone calls and co-authorship. Some links can be observed, but one is attempting to predict unobserved links, or monitor the temporal

aspect; for example, if a snapshot of the set of links at time t is observed then the goal is to predict the links at time $t + 1$.

This problem is normally expressed in terms of a simple binary classification problem. Given two potentially linked objects O_i and O_j , the task is to predict whether L_{ij} is 1 or 0. One approach bases the prediction on the structural properties of the network, for example using predictors based on different graph proximity measures Liben-Nowell and Kleinberg (2003). The second approach is to use attribute information to predict a link. Popescul et al. (2003) applied a structured logistic regression model on relational features to predict the existence of links. A conditional probability model is proposed which is based on attribute and structural features by O'Madadhain et al (2005), (Getoor, 2003; O'Madadhain, 2005; Rattigan & Jensen, 2005). They explain that building statistical models for edge prediction is a challenging problem because the prior probability of a link can be quite small, this makes it difficult to evaluate the model and, more importantly, measure the level of confidence in the predictions. Rattigan and Jensen (2005) propose improving the quality of the predictions by making the predictions collectively. Hence, a number of probabilistic approaches have been developed, some network structure models are based on the Markov Random Field (MRF) model (Chellappa & Jain, 1993) others on Relational Markov Network (Taskar et al., 2003) and, more recently, the Markov Logic Network (Domingos & Richardson, 2004). If case, O represents a set of objects, with X attributes, and E edges among the objects, then MRF uses a joint distribution over the set of edges E , $P(E)$, or a distribution conditioned on the attributes of the nodes, $P(E/X)$. Getoor et al (2003) described several approaches for handling link uncertainty in probabilistic relational models. The key feature of these approaches is their ability to perform probabilistic inferences about the links, which allows the capture of the correlations among the links. This approach is also used for other tasks, such as link-based classification, which allow for more accurate predictions. Hence, approximate inference techniques are necessary to join the model-based probabilistic approaches based on their computational cost to exact inference as general intractable goals.

Desjardins and Gaston (2006) discuss the relationship between the fields of statistical relational learning (SRL) and multi-agent systems (MAS) using link prediction methods to recognise collusion among agents, and applying graph classification to discover efficient networks for MAS problems. Mustafa et al. (2007) show a general approach for combining

object classification and link prediction using Iterative Collective Classification and Link Prediction (ICCLP) in graphs.

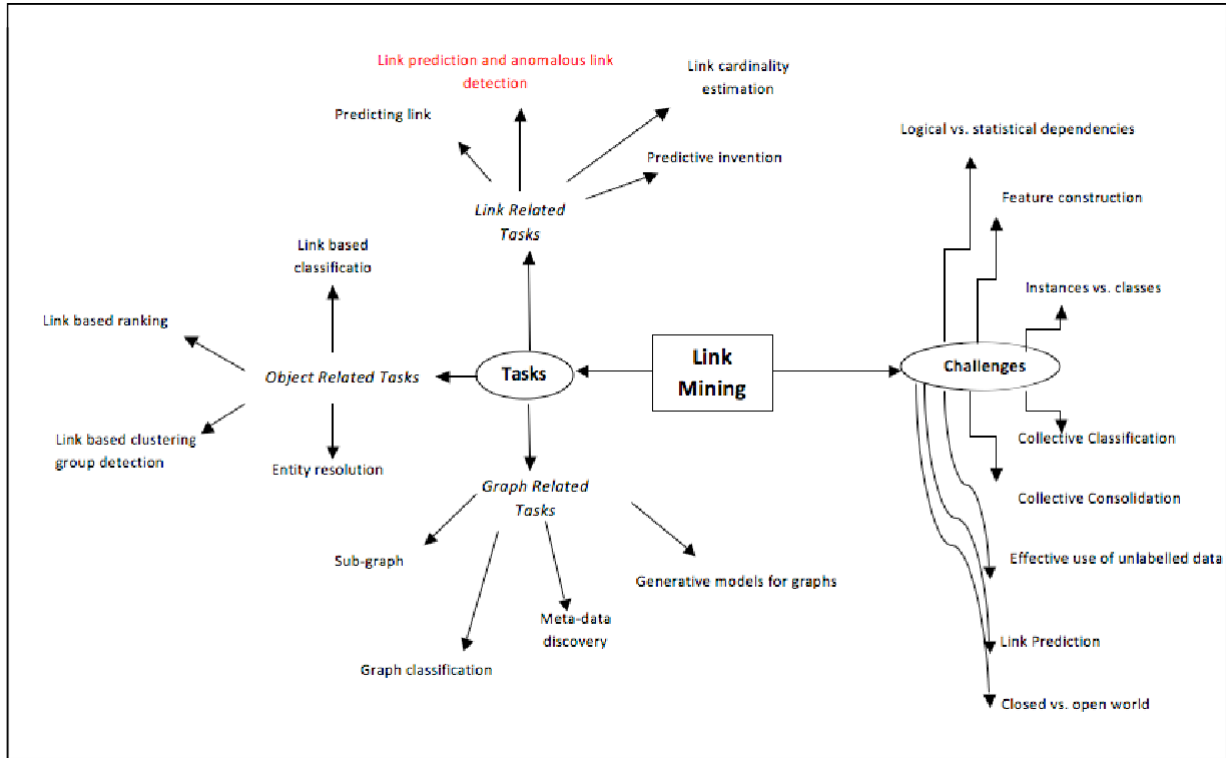


Figure 2. 1 Link mining tasks and challenges

2.3 Link mining challenges

Research into link mining involves a set of challenges associated with these tasks, as Senator (2005), Getoor (2005) and Pedreschi (2008) explain (see Figure 2.1). These are:

- logical vs statistical dependencies that relate to the identification of logical relationships between objects and statistical relationships between the attributes of objects;
- feature construction, which refers to the potential use of the attributes of linked objects;
- collective classification using a learned link-based model that specifies a distribution over link and content attributes, which may be correlated through these links;
- effective use of unlabelled data using semi-supervised learning, co-training and transductive inference to improve classification performance;

- link prediction, which predicts the existence of links between objects;
- object identity, that is, determining whether two objects refer to the same entity; and closed world vs open world assumptions of whether we know all the potential entities in the domain.
- the challenge of this study is to identify and interpret anomalies among the observed links.

2.4 Applications of link mining

An application for each of the three tasks is listed below.

- Social bookmarking is an application of a link-related task. Tools enable users to save URLs for upcoming reference, to create labels for annotating web pages, and to share web pages they found interesting with others. The application of link mining to social web bookmarking investigates user bookmarking and tagging behaviours, and describes several approaches to finding patterns in the data (Chen & Pang-Ning, 2009).
- Epidemiological studies are an application associated with object-related task. In an epidemiology domain, the objects include patients, people with whom they have come into contact and disease strains. Links represent contacts between people and a disease strain with which a person is infected (Getoor, 2003).
- Friendship in a social network is an application of graph-related task. This is annotated by the inclusion of the friend's name on a user's homepage. Pair-dependent descriptions, such as the size of the intersection of interests, offer supplementary evidence for the existence of a friendship. These pair-dependent features are used to determine the probability for link existence where it is not annotated. Finding the non-obvious pair-dependent features can be quite difficult as it, requires the use of recent developments in association rule mining and frequent pattern mining to find correlations between data points that best suggest link existence (Han *et al.*, 2001).
- Bibliographic area is an application of a graph-related task. Information networks are mainly new. Link information in a bibliographic database provides in-depth information about research, such as the clustering of conferences shared by many common authors, the reputation of a conference for its productive authors, research evolving with time, and

the profile of a conference, an author, or a research area. This motivates the study of information network in link mining on bibliographic databases (Getoor, 2003).

- Discovery of a fundamental organisation is an application of graph-related task. Structure from crime data leads the investigation to terrorist cells or organised crime groups, detecting covert networks that are important to crime investigation. (Marcus *et al.*, 2007).

2.5 Anomalies detection

Link prediction is a complex and challenging task as many applications contain data which are extremely noisy and often the characteristics to be employed for prediction are either not readily available or involve complex relationships among objects. The focus of this thesis is to investigate the links between objects and understand the context of their anomalies. Anomaly detection is different from noisy data, which is not of interest to the analyst, and must be removed before any data analysis can be performed. In our research anomalous objects or links can convey useful information and should be investigated.

Song *et al.* (2007) and Chandola *et al.* (2009) describe five types of anomalies, these are:

- Contextual anomalies (also known as conditional anomalies) refer to data instances anomalous in a specific context. A temperature of 5°C might be normal during the winter period in the UK, but would be an anomaly in the summer time.
- Point anomalies refer to a data instance anomalous with respect to the rest of the data set. In credit card fraud application, a transaction is considered a point anomaly if it contains a very high amount spent compared to the normal range of expenditure for that individual.
- Collective anomalies refer to a set of data instances anomalous with respect to the entire data set. For example an electrocardiogram output may show a region of low values for an abnormally long time due to some premature contractions (Goldberger *et al.*, 2002). These low values may not be anomalies by themselves, but their existence together as a collection is anomalous.
- On-line anomalies refer to data present often in a streaming mode where the normal behaviour is changing dynamically.
- Distributed anomalies refer to detecting anomalies in complex systems.

The definition of anomaly is dependent on the type of application domains. For example, in the medical domain a small deviation from normal (e.g., fluctuations in body temperature) could be an anomaly, however similar deviation in the stock market domain (e.g., fluctuations in the value of a stock) might be considered as normal. Thus applying a technique developed in one domain to another has to take into consideration the context of that domain.

Anomalies detection is alike to link prediction in the sense that they both use similar metrics to evaluate which links are anomalous and which ones are expected. Thus research on improving either problem should benefit the other. Rattigan and Jensen explain that one of the important challenges in link prediction is to address the problem of a highly skewed class distribution caused by the fact that “... as networks grow and evolve, the number of negative examples (disconnected pairs of objects) increases quadratically while the number of positive examples often grows only linearly” (Rattigan and Jenssen 2005: 41). As a result, evaluating a link prediction model becomes a complex task and computationally costly because of the need to evaluate all potential links between all pairs of objects. They have proposed the alternative task of anomalous link discovery (ALD) focusing on those links that are anomalous, statistically unlikely, and most “interesting” links in the data. Typical applications of anomaly detection algorithms are employed in domains that deal with security and privacy issues, terrorism activities, picking intrusion detection and illegitimate financial transactions (See Figure 2.1).

2.6 Anomalies detection approaches and methods

A survey of the literature reveals three main approaches used to detect anomalies. These are described below:

- *Supervised* anomalies detection operates in supervised mode and assumes the availability of a training data set, which has labels available for both normal and anomalous data. Typical approach in such cases is to build a predictive model for normal vs. anomalous classes; their disadvantage is that they require labels for both normal and anomalous behaviour. Certain techniques insert artificial anomalies in a normal data set to obtain a fully labelled training data set and then apply supervised anomalies detection techniques to detect anomalies in test data (Abe *et al.*, 2006).

- *Semi-supervised* anomalies detection, which models only normality and are more applicable than the previous approach since only labels for normal data is required. Such techniques are not used commonly, as it is difficult to obtain a training data set which covers possible outlying behaviour that can occur in the data (Chandola *et al.*, 2009).
- *Unsupervised* anomalies detection, which makes the implicit assumption that normal instances are more frequent than anomalies in the test data. If this assumption is not true then such techniques suffer from a high false alarm rate (Chandola *et al.*, 2009).

Unsupervised method is very useful for two reasons. First, they do not rely on the availability of expensive and difficult to obtain data labels; second, they do not assume any specific characteristics of the anomalies. In many cases, it is important to detect unexpected or unexplained behaviour that cannot be pre-specified. Since the unsupervised approach relies on detecting any observation that deviates from the normal data cases, it is not restricted to any particular type of anomaly.

In their paper, Chandola *et al.* (2009) identify five different methods employed in anomalies detection: nearest neighbour, clustering, statistical, classification, and information/ context based approaches (see Figure 2.2).

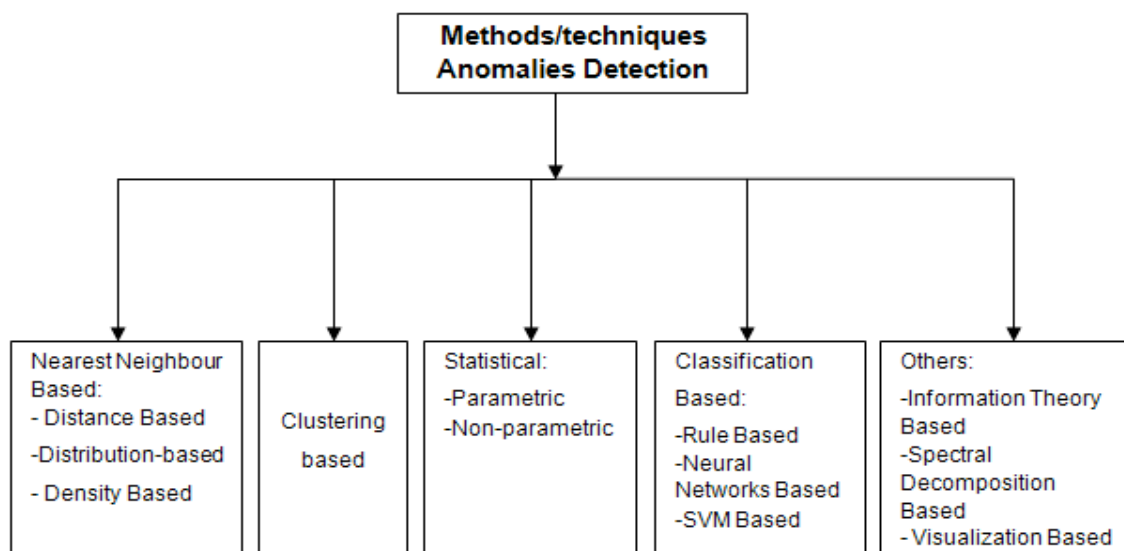


Figure 2. 2 methods of anomalies detection

2.6.1 Nearest neighbour based detection techniques

The concept of nearest neighbour has been used in several anomaly detection techniques. Such techniques are based on the following key assumption:

Assumption: Normal data instances happen in dense neighbourhoods, while anomalies occur far from their closest neighbours.

The nearest neighbour based method can be divided into three main categories. The first distance-based methods, distinguish potential anomalies from others based on the number of objects in the neighbourhood (Hu and Sung, 2003). The distribution-based approach deals with statistical methods that are based on the probabilistic data model, which can be either a automatically or priori, created using given data. If the object does not suit the probabilistic model, it is considered to be an outlier (Petrovskiy, 2003). The density-based approach detects local anomalies based on the local density of an object's neighbourhood (Jin *et al.*, 2001). A typical application area is fraud detection (Ertoz *et al.*, 2004; Chandola et al. 2006), Eskin *et al* (2002).

Nearest neighbour based techniques have many advantages. Key advantage is that they are unsupervised in nature and do not make any assumptions concerning the generative distribution of the data. Instead, it is purely data driven. Adapting these techniques to a variety of data type requires defining a distance measure for the given data. With regards to mixed anomalies, semi-supervised techniques perform more improved than unsupervised techniques since the likelihood of an anomaly is to form a near neighbourhood when the training data set is low.

However, these techniques have disadvantages. They fail to label the anomalies correctly, resulting in missed anomalies, for unsupervised techniques. If the data has normal instances that do not have close neighbours or if the data has anomalies that have close neighbours the technique fails to label them correctly, resulting in missed anomalies. The computational complexity of the testing phase is a challenge since it involves computing the distance of each test instance with all instances belonging to either the test data itself, or to the training data. In semi-supervised techniques, if the normal instances in the test data do not have enough similar normal instances in the training data, then the technique will have a high false positive rate.

2.6.2 Clustering-based anomalies detection techniques

Clustering-based anomalies detection techniques can be grouped into three assumptions:

The first assumption: *Normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster.* Techniques based on this assumption apply a known clustering-based algorithm to the data set and declare any data instance that does not belong to any cluster as anomalous. Several clustering algorithms do not force every data instance to belong to a cluster, such as *DBSCAN* (Ester *et al.*, 1996), *ROCK* (Guha *et al.*, 2001) and *SNN clustering* (Ert  oz *et al.*, 2003). The *FindOut* algorithm (Yu *et al.*, 2002) is an extension of the *WaveCluster* algorithm (Sheik-holeslami *et al.*, 1998) in which the detected clusters are removed from the data and the residual instances are declared as anomalies. A disadvantage of these techniques is that they are not optimised to find anomalies, as the main aim of the underlying clustering algorithm is to find clusters. Typical application areas include image processing (Scarth *et al.*, 1995), and fraud detection (Wu and Zhang, 2003; Otey *et al.* 2003).

The second assumption: *Normal data instances lie close to their closest cluster centroid, while anomalies are far away from their closest cluster centroid.* Techniques based on this assumption consist of two steps. In the first step, the data is clustered using a clustering algorithm. In the second step, for each data instance, its distance to its closest cluster centroid is calculated as its anomaly score. A number of anomaly detection techniques that follow this two-step approach have been proposed using different clustering algorithms. Smith *et al.* (2002) study *Self-Organizing Maps (SOM)*, *K-means* and *Expectation Maximization (EM)* to cluster training data and then use the clusters to classify test data. In particular, SOM (Kohonen, 1997) has been widely used to detect anomalies in a semi-supervised mode in several applications such as intrusion detection (Labib and Vemuri, 2002; Smith *et al.*, 2002; Ramadas *et al.*, 2003), fault detection (Harris, 1993; Ypma & Duin, 1998; Emamian *et al.*, 2000) and fraud detection (Brockett *et al.*, 1998). Barbara *et al.* (2003) propose a robust technique to detect anomalies in the training data. This assumption can also operate in a semi-supervised mode, in which the training data are clustered, with instances belonging to the test data being compared against the clusters to obtain an anomaly score for the test data instance (Marchette, 1999; Wu and Zhang, 2003; Vinueza and Grudic, 2004; Allan *et al.*, 1998). If the

training data have instances belonging to multiple classes, semi-supervised clustering can be applied to improve the clusters to address this issue.

The third assumption: *Normal data instances belong to large and dense clusters, while anomalies belong either too small or too sparse clusters.* Techniques based on the above assumption declare instances belonging to cluster as anomalous if size/density is below a threshold. Several variations of the third assumption of techniques have been proposed (Pires and Santos-Pereira, 2005; Otey *et al.*, 2003; Eskin *et al.*, 2002; Mahoney *et al.*, 2003; Jiang *et al.*, 2001; He *et al.*, 2003). The technique proposed by He *et al.* (2003), called *FindCBLOF*, assigns an anomaly score known as the Cluster-Based Local Outlier Factor (CBLOF) to each data instance. The CBLOF score captures the size of the cluster to which the data instance belongs, in addition to the distance of the data instance to its cluster centroid. These techniques are used for network intrusion detection (Bolton & Hand 1999), and for host based intrusion detection (Sequeira & Zaki 2002).

In terms of advantages these techniques can work in an unsupervised mode, and can be adapted to complex data types by working in a clustering algorithm that can handle the specific data type. The testing stage for clustering based techniques is fast because the number of clusters against is a small constant. However these techniques are highly dependent on the effectiveness in capturing the cluster structure of normal instances. Numerous techniques detect anomalies as a result of clustering, and are not improved for anomaly detection. Some clustering algorithms are assigned to a particular cluster. This could result in anomalies getting assigned to a larger cluster, thus being considered as normal instances by techniques that work under the assumption that anomalies are not linked to any cluster. If $O(N^2d)$ clustering algorithms are used, then the computational complexity for clustering the data is often a bottleneck.

2.6.3 Statistical techniques

Statistical anomaly detection techniques are based on the following key assumption: **Assumption:** Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model.

Statistical techniques operate in two phases: *training* and *testing* phases, once the probabilistic model is known. In the *training* phase, the first step comprises fitting a statistical

model to the given data, whereas the *testing* phase, determines whether a given data instance is anomalous with respect to the model or not. This involves computing the probability of the test instance to be generated by the learnt model. Both parametric and non-parametric techniques are used. Parametric techniques assume the knowledge of underlying distribution and estimate the parameters from the given data (Eskin 2000). Non-parametric techniques do not assume any knowledge of distribution characteristics (Desforges *et al.*, 1998). Typically the modelling techniques are robust to small amounts of anomalies in the data and hence can work in an unsupervised mode. Statistical techniques can operate in unsupervised settings, semi-supervised and supervised settings. Supervised techniques estimate the probability density for normal instances and outliers. The semi-supervised techniques estimate the probability density for either normal instances, or anomalies, depending on the availability of labels. Unsupervised techniques define a statistical model, which fits the majority of the observations. One such approach is to find the distance of the data instance from the estimated mean and declare any point above a threshold to be anomalies (Grubbs 1969). This requires a threshold parameter to determine the length of the tail, which has to be considered as anomalies; techniques used for mobile phone fraud detection (Cox *et al.*, 1997).

The advantages of these techniques are as follows:

- If the assumptions concerning the underlying data distribution are true, these techniques then offer a statistically correct solution for anomaly detection.
- Confidence interval is associated with the anomaly score provided by a statistical technique, which can be used as extra information when making a decision concerning any test instance.
- It can operate in an unsupervised setting without any need for labelled training data if the distribution estimation step is robust to anomalies in data.

However, they rely on the assumption that the data is conducted from a particular distribution. This assumption is not necessarily true, particularly for high dimensional real data sets. Even when the statistical assumption can be justified, there are several hypothesis test statistics that can be useful to detect anomalies; choosing the greatest statistic is often not an easy task (Motulsky, 1995). In specific, composing hypothesis tests for complex distributions needed to fit high dimensional data sets is nontrivial. An anomaly might have attribute values that are individually very common, but their combination is very uncommon,

but an attribute-wise histogram based technique would not be able to detect such anomalies. Histogram based techniques are relatively simple to apply, a key disadvantage of such techniques with regards to multivariate data is that they are not able to capture the interactions between different attributes.

2.6.4 Classification techniques

Classification based techniques operate under the following general assumption:

Assumption: A classifier that can distinguish between normal and anomalous classes can be learnt in the given feature space.

Classification is an important data-mining concept. The aim of classification is to learn a set of labelled data instances (training) and then classify an unseen instance into one of the learnt class (testing). Anomalies detection techniques based on classification also operate in the same two-phase, using normal and anomalies as the two classes. The training phase builds a classification model using the available labelled training data. The testing stage classifies a test instance using the model learnt. The techniques following this approach fall under supervised anomalies detection techniques. A one-class classifier can then be trained to reject this object and to label it as anomalies. These techniques fall under the category of semi-supervised anomalies detection techniques (Tan *et al.* 2005b; Duda *et al.* 2000).

The classification problem is modelled as a two-class problem where any new instance that does not belong to the learnt class is anomalous. In real scenarios, class labels for normal class are more readily available but there are also cases where only anomalies class labels are available. Classification based techniques are categorised into subcategories based on the type of classification model that use. These include Neural networks, Bayesian Networks, Support Vector Machines (SVM), decision trees and regression models. These rules are used to classify a new observation as normal or anomalous.

In term of advantages, the testing stage of these techniques is fast since each test instance needs to be compared against the pre-computed model. They can make use of powerful algorithms that can differentiate between instances belonging to different classes. However, Multi-class classification techniques rely on availability of precise labels for different normal classes, which is often not possible. These techniques allocate a label to each test instance, which can become a disadvantage when a meaningful anomaly score is wanted for the test

instances. Some classification techniques that obtain a probabilistic prediction score from the output of a classifier can be used to address this issue (Platt, 2000).

2.6.5 Information Theory Based

These techniques are based on the following key assumption:

Assumption: Anomalies in data induce irregularities in the information content of the data set.

Information theory based techniques analyse the information content of a dataset using different information theoretic measures such as relative entropy, entropy, *etc.* The general idea is that normal data is regular in terms of a certain information theoretic measure. Anomalies significantly change the information content of the data because of their surprising nature. Thus, the typical approach adopted by this technique is to detect data instances that induce irregularity in the data, where the regularity is measured using a particular information theoretic measure. Information theory based techniques operate in an unsupervised mode.

The advantages of these techniques are as follows:

- They can function in an unsupervised setting.
- They make no assumptions regarding the underlying statistical distribution of the data.

However, the performance of these techniques is greatly dependent on the choice of the information theoretic measure. Frequently, these measures can detect anomalies only when there are large numbers of anomalies existing in the data. It is often nontrivial to obtain when these techniques are applied to sequences and spatial data sets because they rely on the size of the substructure. Another disadvantage is that it is difficult to associate an anomaly score with a test instance using these techniques.

2.6.6 Other Techniques

These techniques are based on the following key assumption:

Assumption: Data can be embedded into a lower dimensional subspace in which normal instances and anomalies appear significantly different.

Spectral decomposition based technique finds an approximation of the data using a combination of attributes that capture the size of variability in the data. The underlying

assumption for such techniques is that the reduced sets of attributes faithfully capture much of the normal data, but this is not necessarily true for the anomalies. Spectral techniques can work in an unsupervised as well as semi-supervised setting. This approach has been applied to the network intrusion detection domain by several different groups (Shyu *et al.* 2003; Lakhina *et al.* 2005; Thottan and Ji 2003) and for detecting anomalies, for example in spacecraft components (Fujimaki *et al.* 2005).

Visualisation based technique maps the data in a coordinate space that makes it easy to visually identify the anomalies. Cox *et al.* (1997) present a visualisation-based technique to detect telecommunications fraud, which displays the call patterns of various users as a directed graph such that a user can visually identify abnormal activity.

These techniques routinely perform dimensionality reduction, which makes them suitable for handling high dimensional data sets. Additionally, they can be used as a pre-processing step, followed by application of any existing anomaly detection technique in the transformed space. These techniques can be used in an unsupervised setting.

However, these techniques usually have high computational complexity. They are useful only if normal and anomalous instances are separate in the lower dimensional embedding of the data.

2.6.7 Overview of strengths and limitations

For high-dimensional data, any of the above anomalies detection techniques can easily detect the anomalies. For more complex data sets, different techniques face different challenges. Chandola *et al.* (2009) argue that statistical techniques do not work well with high-dimensional categorical data and that visualisation-based techniques are more naturally suited to low-dimensional data and hence require dimensionality reduction as a pre-processing step when dealing with a higher number of dimensions. Spectral decomposition-based techniques, which find an approximation of the data using a combination of attributes to capture the variability in the data, explicitly address the high-dimensionality problem by mapping data to a lower dimensional projection, but their performance is highly dependent on the fact that the normal instances and anomalies are distinguishable in the projected space. Clustering is often called an unsupervised learning task, as no class values indicate an a priori grouping of the data instances, as in the case for supervised learning. Clustering and nearest neighbour techniques rely on a good similarity or distance measure to handle the anomalies in complex

data sets. Classification-based techniques handle the dimensionality better, since they try to assign weights to each dimension and ignore unnecessary dimensions automatically. However, classification-based techniques require labels for both normal data and anomalies. Finally, information theory-based techniques, which analyse the information content of a data set using different information theoretic measures (e.g. entropy measure), require a measure that is sensitive enough to detect the effects of even single anomalies. Such techniques detect anomalies only when there is a significant number of an anomaly.

2.7 Challenges of anomalies detection

Multi- and high-dimensional data make the outlier mining problem more complex because of the impact of the curse of dimensionality on algorithms' performance and effectiveness. Wei *et al.*, (2003) introduce an anomalies mining method based on a hyper-graph model to detect anomalies in a categorical data set. He *et al.* (2005) define the problem of anomalies detection in categorical data as an optimisation problem from a global viewpoint, and present a local search heuristic-based algorithm for efficiently finding feasible solutions. He *et al.* (2005) also present a new method for detecting anomalies by discovering frequent patterns (or frequent item sets) within the data set. The anomalies are defined as the data transactions that contain less frequent patterns in their item sets. The recent surveys on the subject (Chandola *et al.*, 2009; Patcha & Park, 2007) note that anomalies detection has traditionally dealt with record or transaction type data sets. They further indicate that most techniques require the entire test data before detecting anomalies, and mention very few online techniques. Indeed, most current algorithms assume that the data set fits in the main memory (Yankov *et al.*, 2007). Both aspects violate the requirement for real-time monitoring data streams. In addition, most approaches focus specifically on intrusion detection (Kuang & Zulkernine, 2008; Xu *et al.*, 2005; Lee & Stolfo, 2000). A comparative study (Chandola *et al.*, 2008) of methods for detecting anomalies in symbolic data shows that there are several techniques for obtaining a symbolic representation from a time series (Lin *et al.*, 2007; Bhattacharyya & Borah, 2004), but all such works seem to apply solely to univariate data (Keogh *et al.*, 2004; Wei *et al.*, 2003). It is a challenging task to detect failures in large dynamic systems because anomalous events may appear rarely and do not have fixed signatures.

2.8 Anomalies detection and link mining

The literature review reveals a growing range of applications in anomalies detection, mostly to data mining and very few applications in link mining. In recent years application of anomalies detection in link mining has gained increasing importance. For example, the paper of Savage *et al* (2014) in online social networks survey's existing computational techniques used to detect irregular or illegal behaviour; other works include detecting fraudulent behaviour of online auctioneers (Chan *et al.*, 2006). Community based anomalies detection in evolutionary networks (Chen *et al.*, 2012), link based approach for bibliometric journal ranking (Su *et al.*, 2013). However, their focus is still on pattern finding rather than link related tasks. Even the work on citation data (Wanjantnle and Keane, 2014, Yang *et al.*, 2011) is used to describe communities or computational techniques and not mining anomalies or predictive links. Thus, much of the work in this area has focused on identifying patterns in behaviour of the data rather than link mining. Anomalies detection in link mining is still a emerging area.

2.9 Summary

Link mining is an emerging area within knowledge discovery focused on mining task relationship by exploiting and explicitly modelling the links among the entities. We have overviewed link mining in terms of object related task, link-based object and group related task. These represent some of the common threads emerging from a variety of fields that are exploring this exciting and rapidly expanding field. However, with the introduction of links, new tasks also come to light: predicting the type of link between two objects, predicting the numbers of links, inferring the existence of a link, and inferring the identity of an object. A review of computational techniques is provided outlining their challenges. Anomaly detection, which is important to use in this research, is also discussed and the current methods and issues highlighted.

These two areas are attracting much interest by researchers from different disciplines (*e.g.* computer science, business, statistics, forensics and social sciences) interested in extracting tacit, hidden, but valuable knowledge from the vast amount of data available worldwide. The emphasis in our study is not on the discovery but the interpretation and semantic value of that discovery. We believe mutual information has a role to play in this semantic analysis.

3 Anomalies in link mining based on mutual information

This chapter introduces the novel approach to anomaly detection in link mining, based on the concept of mutual information. The chapter is organised into three parts. The first part introduces the basic concepts of mutual information, followed by a review of major applications of mutual information in anomaly detection and link mining. The second part describes the novel approach of anomaly detection based on mutual information developed to address the gap of using mutual information to detect anomalies in link mining. The third part is to apply this approach in link mining. This research has adapted CRISP data mining methodology to the emerging field of link mining.

3.1 Mutual Information in Information Theory

Information theory is the branch of mathematics that describes how uncertainty should be manipulated, quantified and represented. Ever since the fundamental premises of information theory were laid down in 1949 by Claude Shannon, it has had far reaching implications for almost every field of science and technology. A measure based on information-theoretic principles will remain relevant for any communication medium. Information-theoretic analysis is an effective tool for data exploration as it provides a model-free way to discover unexpected relationships in data (Steeg & Galstyan, 2013). Mutual information can be defined as the amount of information one random variable contains about another. Mutual information is essentially the measure of how much ‘knowledge’ one can gain of a certain variable by knowing the value of another variable. It measures the relevance among data objects under the problem setting. This function is utilised to capture the relations among data objects, whereby the entire objects are represented as an edge-weighted graph where pairs of objects are connected with edges with their relevance.

Mutual information represents the average amount of information about X that can be gained by observing Y ; it measures the amount of reduction of uncertainty in X after Y is known. It is denoted as $I(X, Y)$ and expressed as follows: $I(X, Y) = H(X) - H(X/Y)$, where $H(X/Y)$

represents the amount of information shared between X and Y , $I(X, Y)$ corresponds to the intersection of the information in X with the information in Y .

Definition (mutual information): The mutual information of two discrete random variables X and Y is defined as (Cover & Thomas, 2006) :

$$I(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (1)$$

It measures the distance between the joint distribution $p(x, y)$ and the product distribution $p(x)p(y)$.

Definition (continuous mutual information): The continuous mutual information between two (continuous) random variables with joint density $f(x, y)$ is defined as (Cover & Thomas, 2006):

$$I(X, Y) = \int_S f(x, y) \log_2 \left(\frac{f(x, y)}{f(x)f(y)} \right) dx dy \quad (2)$$

Where S is the support set of $f(x, y)$.

3.1.1 Estimation of mutual information

To estimate the mutual information:

- Let $nb_{x,y}$ be the number of data points such that the random variable X is equal to x and the random variable Y is equal to y .
- Similarly, let $nb_x(nb_y)$ be the number of data points such that $X = x$ ($Y = y$) and let n be the total number of data points.

Definition (estimated MI): The estimated MI of two random variables X and Y is defined by (Cover & Thomas, 2006):

$$\hat{I}(X, Y) = \sum_{x,y} \sum_{y \in \mathcal{Y}} \hat{p}(x, y) \log_2 \left(\frac{\hat{p}(x, y)}{\hat{p}(x)\hat{p}(y)} \right) \quad (3)$$

$$= H^{\wedge}(X) + H^{\wedge}(Y) - H^{\wedge}(X, Y) \quad (4)$$

$$\text{where } p^{\wedge}(x) = \frac{nb_x}{n}, \quad p^{\wedge}(y) = \frac{nb_y}{n} \text{ and } p^{\wedge}(x, y) = \frac{nb_{x,y}}{n}.$$

The two random variables X and Y are independent if and only if $I(X, Y) = 0$. This fact can then be used to estimate the dependency between X and Y . The simplest way to do this is to define a threshold I_0 and simply say that X and Y are dependent if $I(X, Y) > I_0$. But the problem with this estimation is that one has to define the threshold I_0 and a priori we have no idea how good this threshold is. What we really would like to have is a statistical test where we can decide whether X and Y are dependent and where we have a confidence level α .

3.1.2 Entropy vs. mutual information

The concept of information is too broad to be captured in one single definition whereas mutual information is a measure of the amount of information one random variable contains about entropy of a random variable as it measures its unpredictability (Miller *et al.*, 2013). Mutual information is a special case of a more general quantity called relative entropy; it is a measure of the distance between two probability distributions.

Entropy is a measure of uncertainty and unpredictability of a random variable in information theory (Cover & Thomas, 2006; Shannon, 1948). The entropy tells how much information there is in an event. Overall, the more random or uncertain the event is, the more information it will contain. Once having defined the entropy of a single random variable, one can then define the joint entropy and the conditional entropy.

Definition (joint entropy): The joint entropy $H(X, Y)$ of two discrete random variables X and Y with the joint distribution $p(x, y)$ is defined as:

$$H(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 \frac{1}{p(x, y)} \quad (5)$$

Definition (conditional entropy): For two discrete random variables X and Y with joint distribution $p(x, y)$ the conditional entropy $H(X/Y)$ is defined as:

$$H(X/Y) = \sum_{y \in \mathcal{Y}} p(y) H(X/Y=y) \quad (6)$$

$$= \sum_{y \in \mathcal{Y}} p(y) \left(- \sum_{x \in \mathcal{X}} p(x/y) \log_2 p(x/y) \right) \quad (7)$$

$$= \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x/y) \quad (8)$$

One can show that the joint entropy is the sum of the marginal and the conditional entropy (Cover and Thomas, 2006):

$$H(X, Y) = H(X) + H(X/Y) \quad (9)$$

The relationship between entropy and mutual information can be captured using a *Venn diagram* (see Figure 3.1). To visualise these asset quantities is reasonable, since they behave like sets.

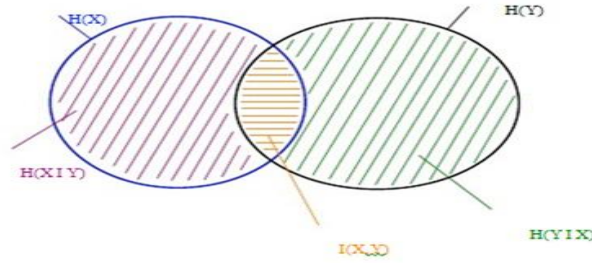


Figure 3. 1 Venn diagram showing entropy, conditional entropy and mutual information

The uncertainty in X is drawn as a circle (blue) and the uncertainty in Y as another circle (black). If the two random variables X and Y were independent, the two circles would not touch each other. If the two entropy circles overlap (see Figure 3.1), the two variables are dependent. If we know variable X , and there is no uncertainty in X then the uncertainty in Y that remains is $(H(X/Y))$ depicted in the green part in Figure 3.1. On the other hand, if we know all about Y , the uncertainty in X that remains is $(H(X/Y))$, the violet part in the above figure. The orange part is the information that both variables share $(I(X, Y))$. The bigger this orange part is, the stronger is the dependency between X and Y . The two variables are independent if, and only if, the orange part $(I(X, Y))$ is zero.

3.1.3 Applications of mutual information

Early applications of mutual information focused on telegraph and radio communications. In telecommunications, the channel capacity is equal to the mutual information, maximised over all input distributions. Networks from different knowledge domains share quite a number of similarities, and researchers have started to analyse networks from different knowledge domains using similar techniques and describe them using similar models (Newman, 2003;

Watts & Strogatz, 1998). Eagle and Pentland's (2006) study on reality mining (Eagle, Pentland & Lazer, 2008, 2009) is perhaps the only study that uses information theory to construct measures of human behaviour.

A survey of the literature review reveals that mutual information is applied in a variety of fields ranging from medical, engineering, search engines, social networks, data and text mining. Some of these applications are described below. Mutual information is used in medical imaging for image registration. Given a reference image (e.g. a brain scan), and a second image, that needs to be put into the same coordinate system as the reference image, this image is deformed until the mutual information between it and the reference image is maximised (Chai et al., 2009). In networks and in bioinformatics, mutual information is commonly used to estimate gene–gene associations based on the expression patterns as represented in sequential lists of nucleotides (Butte & Kohane, 2000; Dawy *et al.*, 2006). Another application of mutual information in bioinformatics is between genes in expression microarray data and is also used by the ARACN, E algorithm for reconstruction of gene networks. Phylogenetic profiling prediction from pairwise present and the disappearance of functionally link genes are used for the prediction of protein structures (Adami, 2004), or boosting and facial expression recognition (Shan et al., 2005). Both entropy and mutual information have been used for independent component and subspace analysis (Learned-Miller and Fisher, 2003; P'oczoz and L'orincz, 2009; Hulle, 2008; Szab'o et al., 2007), and image registration (Kybic, 2006; Hero et al., 2002b, a). These are based on the idea that entropy and mutual information are determined solely by the density. Mutual information is used to discover functional linkages (Date & Marcotte. 2003). It is used as a phylogenetic profiling of proteins and as a metric to cluster proteins based on their profiles. Further applications of mutual information include Bindewald and Shapiro (2006) who used mutual information between positions on sequence alignments as a feature for the prediction of RNA secondary structure. Tomovic and Oakeley (2007) also used mutual information in transcription factor binding site analysis to identify highly correlated positions. Buslje et al. (2010) have shown that networks of high mutual information define the structural proximity of catalytic sites and can be used for their prediction. Finally Brunel et al (2010) have devised a 'mutual information statistical significance' test for genetic association studies.

Mutual information is used to learn the structure of Bayesian Networks. In text mining, computational linguistics researchers have developed algorithms to calculate word

associations based on their occurrences in a large corpus of text documents (for example, Church & Hanks, 1990; P. Li & Church, 2007; Seretan & Wehrli, 2006). Mutual information of words is often used as a significance function for the computation of collocations in corpus linguistics. This has the added complexity that no word-instance is an instance to two different words; rather, one count instances where 2 words occur adjacent or in close proximity; this slightly complicates the calculation, since the expected probability of one word occurring within M words of another, goes up with M.

In bioinformatics, mutual information is used to group together genes with similar patterns of expression (Eisen et al., 1998). The result of this study (Steuer, 2002) was exemplified using a publicly available dataset corresponding to up to 300 diverse mutations and chemical treatments in *S. cerevisiae* (Hughes et al., 2000). The detection of relationships between two or more variables is not restricted to the analysis of gene expression, but is of great importance in various areas of science. Variables which are not statistically independent suggest the existence of some functional relation between them. While there are several approaches to quantify the linear dependence between variables, the framework of information theory (Shannon, 1948) provides a general measure of dependencies between variables. In particular, a disappearing Pearson correlation does not imply that two variables are independent. The mutual information therefore provides a better and more general criterion to investigate relationships between variables.

With the application of data mining that increases data dimensionality in many domains such as bioinformatics, text categorisation, and image recognition, feature selection has become an important data mining preprocessing methods. Mutual information has been used in feature selection (Peng and Ding, 2005), clustering (Aghagolzadeh et al., 2007), causality detection (Hlaváckova-Schindler et al., 2007), and optimal experimental design (Lewi et al., 2007; Póczos & L'orincz, 2009). The aim of feature selection is to find a minimal feature subset of the original datasets that is the most characterising. Dash & Liu, (1997) point out that there are four basic steps in a typical feature selection method that is; subset evaluation, subset generation, stopping criterion, and validation. Zilin et al (2014) propose a new algorithm that combined rough conditional entropy and a naive Bayesian classifier to select features.

3.2 Proposed novel approach

The problem of detecting anomalies has been studied, in particular, from a statistical perspective. Statistical distribution is applied to model data points, which are analysed to determine whether they are to be anomalies in relation to the model. The main problem with such an approach is that, in a number of cases, it might not have enough knowledge about the underlying data distribution (Ramaswamy *et al.*, 2000). Anomalies can be removed or considered separately in regression modelling to improve accuracy, which can be considered a benefit of anomalies. Identifying them prior to modelling and analysis is important (Williams *et al.*, 2002).

However, anomalies in data can reveal significant information in many applications and link mining in particular. The proposed study advocates the use of mutual information to study the relationships between anomalies objects/entities. Based on information theory, mutual information provides a general measure of dependencies between variables. The proposed approach is novel as it uses mutual information to analyse anomalies in data sets and investigates the semantic interpretation of the link that relates one object to another. The novel method is applied to two new areas: transaction data, and citation data.

3.3 Methodology of link mining

The field of data mining over the past few years is becoming extremely important for businesses, co-operations, companies and industries *etc.*, Different process models were introduced to the field of data mining to carry and guide data mining applications and tasks. The three most popular data mining process models are Knowledge Discovery Databases (KDD) process model, CRISP-DM and SEMMA (Shafique & Qaiser 2014). The Knowledge Discovery Databases (KDD) process model is interactive (Brachman & Anand 1996); it consists of nine steps and emphasise database, as it is primary data source. Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman 2000) was first launched in 1996 by Daimler, then improved and refined over the years, it consists of six phases. SEMMA developed by SAS Enterprise Miner institute (2014) has five phases: Sample, Explore, Modify, Model, and Assess. In this study CRISP-DM has been adapted to the field of link mining.

3.3.1 Knowledge Discovery Databases (KDD)

KDD (Knowledge Discovery Databases) is the process of extracting the hidden knowledge from data (Fayyad et al., 1996). There are nine different steps or stages:

1. Understanding the application domain: This is the first stage of KDD in which goals are defined, used to develop an understanding about application domain and its prior knowledge
2. Creating a target data set: The second stage of KDD focuses on creating target data set and subset of data variables. It is an essential step as knowledge discovery is performed on all these.
3. Data cleaning and data pre-processing: This is the third stage of KDD focuses on data cleaning and pre-processing to complete data without any noise. In this stage, strategies are developed to handle such type of inconsistent and noisy data.
4. Data transformation: The fourth stage of KDD, focuses on transformation of data from one form to another form enabling data mining algorithms to be easily implemented. For this purpose different data transformation and reduction methods are implemented on target data.
5. Choosing data mining task: This is the fifth stage of KDD where appropriate data mining task is chosen based on particular goals that are stated in the first stage. Examples of data mining tasks are classification, clustering, regression and summarisation, etc.
6. Choosing data mining algorithm: This is the sixth stage of KDD in which appropriate data mining algorithms are chosen for searching different patterns from data. There are many algorithms available today for data mining but suitable algorithms are chosen based on matching the overall criteria for data mining.
7. Employing data mining algorithm: This is the seventh step of KDD in which selected algorithms are implemented.
8. Interpreting mined patterns: This is the eighth stage of KDD that focuses on evaluation and interpretation of mining patterns. This step may involve in visualising extracted patterns.
9. Using discovered knowledge: This is the final stage of KDD in which the discovered knowledge is used for different purposes. The knowledge discovered can be used by interested parties or can be integrated with another system for further actions.

3.3.2 SEMMA

SEMMA (Sample, Explore, Modify, Model, Assess) refers to the process of conducting a DM project (Santos & Azevedo, 2005). It is a data mining method developed by SAS institute. It allows development, maintenance, organisation, and understanding of data mining. It focuses on the model development aspect of data mining. SEMMA is linked to SAS enterprise miner and it is considered more of a functional tool for them rather than a data mining methodology. The process contains five stages:

1. Sample: this stage consists of sampling the data by extracting a sample of a large data set, big enough to contain the significant information, yet small enough to manipulate.
2. Explore: this stage relies on the exploration of the data by searching for unexpected trends and irregularities in order to gain ideas and understanding.
3. Modify: this stage consists on the adjustment of the data through creating, selecting, and transforming the variables to focus the model selection process.
4. Model: this stage consists on modelling the data by allowing the software to search automatically for a combination of data that predicts a desired outcome.
5. Assess: this stage consists on assessing the data by assessing the usefulness of the findings from the DM process and estimating how well it performs. SEMMA offers an easy to understand process, permitting an organised and sufficient development and maintenance of DM projects.

3.3.3 CRISP-DM

This methodology consists of six stages (Figure 3. 2):

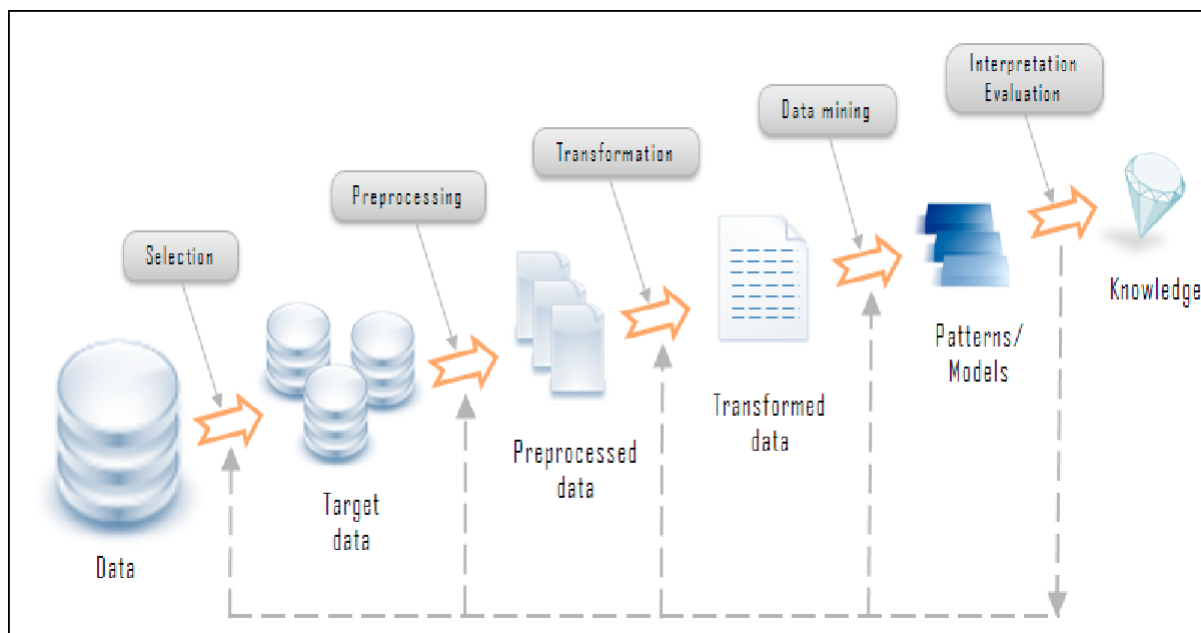


Figure 3. 2 CRISP-Data mining methodology (Shearer 2000)

Stage 1: Problem definition

This stage establishes data mining goals: the objectives are clearly identified and a project plan is developed.

Stage 2: Data understanding

This stage starts with the initial data collection from available data sources. Activities such as initial data collection, data description, and data integration are essential in order to make the data collection successful.

Stage 3: Data pre-processing

Once the data resources available are identified, they need to be selected, cleaned, and formatted/converted appropriately before further exploration.

Stage 4: Data exploration

Data exploration task may be carried out at a greater depth during this phase to identify the patterns in data. Such as, viewing the summary statistics (which includes the visual display of categorical variables) that can occur at the end of this phase. During this phase models such as cluster analysis can also be applied, with the intent of identifying patterns in the data.

Stage 5: Data modelling

In this phase, various modelling techniques are selected and applied, and their factors are adjusted to optimum values. There are several techniques for the same data mining problem type. Some techniques have particular requirements on the form of data, such as visualisation (plotting data and establishing relationships) and cluster analysis (to identify which variables go well together) are useful for initial analysis; more detailed models appropriate to the data type can be applied.

Stage 6: Evaluation and deployment

Evaluation is an integral part of the model development process. It helps find the model that best represents the data and predicts how well the chosen model will work in the future. If the model achieves the objectives defined in stage 1 then a plan of action is developed to apply this model. Before continuing to the final deployment of the model, it needs to undergo a more thorough evaluation, and the steps executed to construct it need to be reviewed, to be certain that it properly achieves the mining objectives. A key objective is to determine important issues that may not have been adequately considered. At the end of this phase, a decision on the use of the data mining results should be reached. In the deployment stage, the creation of the model is not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be presented and organised in a way that can be used. This will lead to the identification of other needs (often through pattern recognition), commonly reverting to prior phases of data mining, where the results of various visualisation, statistical, and artificial intelligence tools show the user new relationships that provide a deeper understanding of organisational operations.

The KDD process (Piatetsky-Shapiro, 1994) has a process model component because it is more accurate, complete and establishes all the steps to be taken to improve a data mining project, but it is not a methodology because its definition does not set out how to perform each of the proposed tasks. It is a generic methodology consisting of nine stages. In contrast, SEMMA is a company¹ oriented approach focusing on SAS Enterprise Miner software and on model development specifically; it places less importance on the initial planning phases, which are covered in CRISP-DM and skips entirely the deployment phase. The SEMMA methodology is only concerned with statistical modelling and practical implementation of the five stages of KDD. It lacks important parts of any information system project including

analysis, design and implementations (Umair & Haseeb 2014). CRISP-DM provides a uniform framework and guidelines for data miners, by working well even with small-scale data mining and different types of data; it is able of discovering hidden anomalous pattern in data (see Table 3.1). We believe that this method can also help provide a structured approach to link mining. In this thesis we have adapted CRISP-DM to link mining in order to find hidden patterns in links and related objects.

Table 3.1. Summary of differences between KDD, CRISP-DM and SEMMA.

Data Mining Process Models	KDD	CRISP-DM	SEMMA
No. of Steps	9	6	5
Name of Steps	Developing and Understanding of the Application	Business Understanding	-----
	Creating a Target Data Set	Data Understanding	Sample
	Data Cleaning and Pre-processing		Explore
	Data Transformation	Data Preparation	Modify
	Choosing the suitable Data Mining Task	Modeling	Model
	Choosing the suitable Data Mining Algorithm		
	Employing Data Mining Algorithm		
	Interpreting Mined Patterns	Evaluation	Assessment
	Using Discovered Knowledge	Deployment	-----

3.4 Link Mining Methodology

As CRISP-DM methodology is well developed and applied in knowledge discovery, this research has adapted it to the emerging field of link mining. While data mining addresses the discovery of patterns in data entities, link mining is interested in finding patterns in objects by exploiting and modelling the link among the objects. The approach to link mining is still an ad-hoc approach. The proposed adopted CRISP-DM methodology can help provide a structured approach to link mining. This consists of six stages:

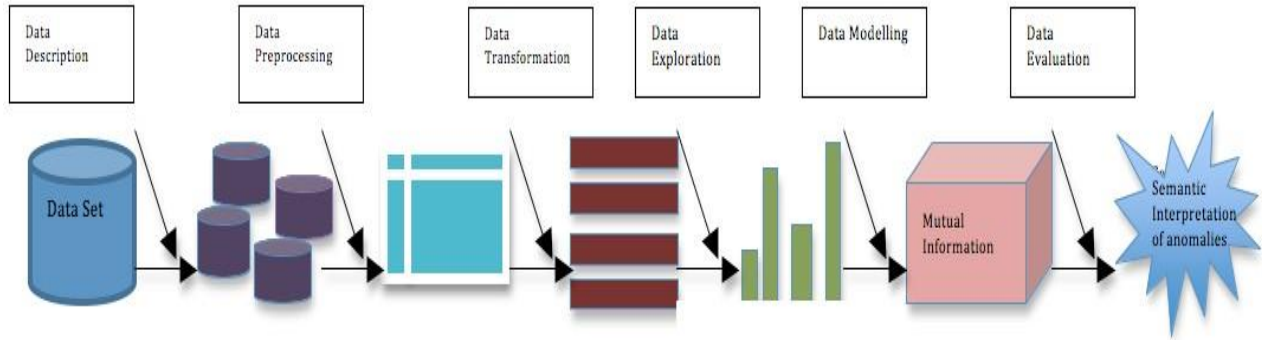


Figure 3. 3 Link mining methodology

The aim of this methodology is to define the link mining task and determining the objectives of link mining.

1. **Data description.** The data description phase starts with initial data collection and proceeds with activities that enable the researcher to become familiar with the data. The aim is to check data quality and any associated problems in order to discover first insights into the data, and identify interesting subsets to form hypotheses regarding hidden information.
2. **Data pre-processing.** The data pre-processing phase covers activities related to data cleansing and data integrity needed to construct the final dataset from the initial raw data. While outliers can be considered noise, or anomalies and thus discarded in data mining, they become the focus of this study as they can reveal important knowledge in link mining.
3. **Data transformation.** This involves syntactic modifications applied to the data; this maybe required by the modelling tool. Selecting an appropriate representation is an important challenge in link mining. The objects in link mining (e.g. people, events, organisation, and countries) have to be transformed into feature factors to represent and capture the connectivity and the strength of the links among those objects.
4. **Data exploration.** This stage is concerned with the distribution of the data and using relevant graphical tools to visualise the structure of the objects and their links. This stage helps identify the existence of anomalous objects or links.
5. **Data modelling.** This stage aims to identify all entities and the relationship between them. Data modelling puts algorithm in general in a historical perspective rooted in mathematics, statistics, and numerical analysis. For more complex data sets, different

techniques are used such as nearest neighbour, statistical, classification, and information/ context based approaches.

- 6 Evaluation: Data cleaning solutions will clean data by cross checking with a validated data set in phase 2. The clustering model in phase 5, explains natural groupings within a dataset based on a set of input variables. The resulting clustering model is sufficient statistics for calculating the cluster group norms and anomaly indices. Mutual information is useful in validating the model as it provides a semantic underpinning to the patterns and discoveries made in phase 5.

3.5 Summary

In the last decade we have seen an increasing interest in the study of anomalies detection in data mining applied to law enforcement, financial fraud, and terrorism. In recent years, this study has been applied to social networks and online communities to identify influential networks participants and predict fraudulent or malicious activities. To our knowledge, the study of anomaly detection in link mining relied mostly on statistical or machine learning methods in order to gain insight to the structure of their networks. We believe that we can achieve a better understanding of these anomalies if we apply mutual information to the data entities and objects and links to reveal their semantic relationship. This chapter introduced the novel approach to anomaly detection in link mining based on mutual information. This proposed novel approach is investigated through the use of two case studies described in the following chapters.

Case study1 is a proof of concept data designed to test the validity of the proposed approach. The aim is to apply the proposed link mining methodology to detect anomalies and identify the sources of these anomalies embedded in this case study. The modelling task will use a two-step clustering setting and apply mutual information between two sets of variables and study their association patterns to estimate the extent to which the two variables co-vary with each other.

Case study 2 is based on the study of real data to demonstrate how mutual information can help explore and interpret anomalies detection with a different data set and application area, such as co citation data, making use of different forms of data representation, for example

graphs to visualise the dataset and applying a different clustering approach (e.g. hierarchical clustering method) in the modelling stage.

4 Anomalies Detection: Case study 1

This chapter investigates the proposed novel anomaly detection method, which advocates the use of a mutual information based measure in order to study the relationships between anomalies and identify vital hidden information in link mining. This method is applied to Case study 1, which is basically a proof of concept consisting of a small data set, constructed with known anomalies.

4.1 Overview of Case study1

Case study 1 is the proof of concept data designed to test the validity of the proposed approach. The aim is to detect anomalies and identify the sources of these anomalies. The data used in this Case study consists of 500 transactions related to purchases undertaken by customers from seven supermarkets (Stafford, Birmingham, Hull, Oxford, Leeds, London and Manchester) on 1st October 2012 during these time slots: pm, am and evening. Each transaction consists of the eleven fields. The analysis focuses on these fields in order to identify any anomalies in the data, to understand their properties, relationships and mine potential links. The tasks are to identify anomalous transactions within data that is seemingly homogeneous and to investigate whether mutual information can help identify the sources of anomalous transactions.

4.2 Anomaly detection methodology applied to Case study 1

The approach taken is based on the extended methodology described in the previous chapter; the 6 stages are applied to the Case study 1 and explained below (see Figure 4.1).

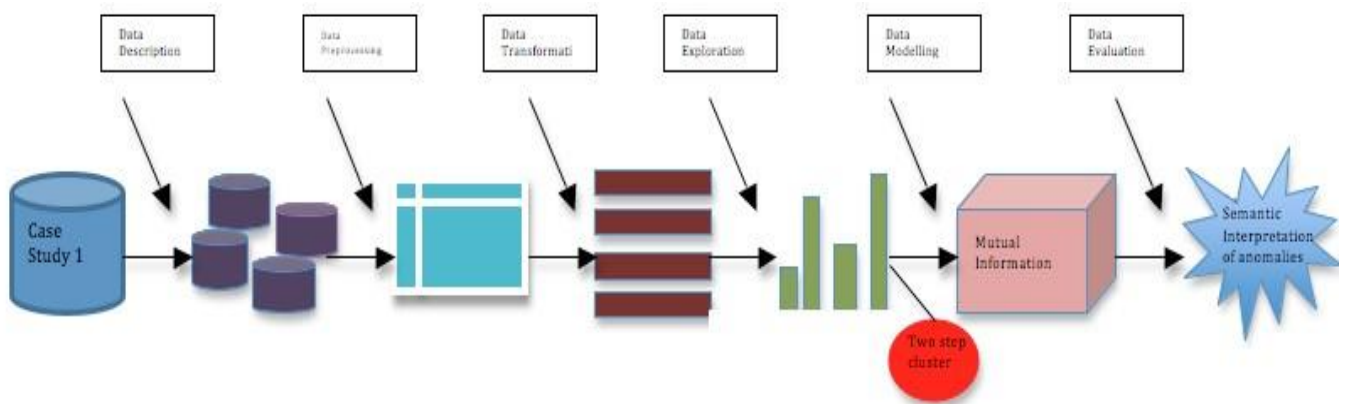


Figure 4. 1 Link mining methodology

The aim of problem definition is to focus on understanding the project objectives and requirements from the domain perspective and then converting this knowledge into a link mining definition with an initial plan designed to achieve the objectives. Errors in the data need to be examined taking into consideration the context of the domains; some may be true errors and therefore removed, whereas other errors may be kept as legitimate anomalies. The challenge of this phase is that in some cases the user may simply not have enough knowledge about the underlying data distribution (Ramaswamy *et al.* 2000), so special care must be given to understand the data context. In Case study 1, the focus is on detecting anomalies in transactions and to identify the source of these anomalies.

4.2.1 Stage 1: Data description

This phase focuses on understanding the properties of the acquired data, and its quality. The data in Case study 1 consists of a set of 500 fictional records related to sales. Each record is related to a particular transaction consisting of 11 fields: purchase category (credit, debit, cash, cheque and vouchers), transaction value ranging from £1 to £1000, sale ID, date, timeslot, stationary product, location, staff ID, month, staff training location and staff trainer' name.

Table 4. 1 A small sample of the case 1 data

SaleID	Date	TimeSlot	StationeryProduct	PurchaseCategory	Location	TransactionValue	StaffID	Month	StaffTrainingLocation	StaffTrainer
A059	01-Oct-12	PM	Pencil	2.00	7.00	56.00	11	April	London	3.00
A060	01-Oct-12	EVE	Pencil Rubber	1.00	7.00	78.00	9	April	London	3.00
A222	01-Oct-12	AM	Stapler Pin	2.00	7.00	78.00	7	July	London	3.00
A246	01-Oct-12	EVE	Pencil	2.00	2.00	1.00	10	Aug	London	5.00
A286	01-Oct-12	EVE	Stapler Remover	5.00	7.00	56.00	13	Aug	London	3.00
A397	01-Oct-12	PM	Correction	1.00	2.00	53.00	7	Sept	London	5.00
A412	01-Oct-12	AM	Pencil	3.00	7.00	1.00	3	Nov	London	3.00
A460	01-Oct-12	AM	Correction	1.00	2.00	20.00	7	Sept	London	1.00
A488	01-Oct-12	AM	Stapler Pin	3.00	5.00	20.00	12	Aug	London	2.00
A482	01-Oct-12	PM	Stapler Remover	3.00	2.00	54.00	6	Dec	London	2.00
A204	01-Oct-12	AM	Gift pen Set	1.00	5.00	27.00	3	July	London	2.00
A001	01-Oct-12	Am	Laser printer	3.00	7.00	37.00	1	Jan	London	4.00
A028	01-Oct-12	AM	Laser printer	4.00	1.00	56.00	4	Feb	London	3.00
A048	01-Oct-12	PM	Laser printer	4.00	1.00	78.00	6	March	London	4.00
A105	01-Oct-12	PM	Laser printer	3.00	1.00	78.00	15	May	London	3.00
A189	01-Oct-12	AM	Laser printer	3.00	1.00	1000.00	12	June	London	3.00
A284	01-Oct-12	PM	Laser printer	1.00	3.00	56.00	9	Aug	London	1.00
A410	01-Oct-12	EVE	Laser printer	4.00	7.00	53.00	2	Sept	London	3.00
A450	01-Oct-12	PM	Laser printer	3.00	7.00	194.00	1	Dec	London	3.00
A477	01-Oct-12	EVE	Laser printer	4.00	6.00	100.00	4	Dec	London	3.00
A497	01-Oct-12	AM	Laser printer	4.00	6.00	1000.00	6	Aug	London	3.00

4.2.2 Stage 2: Data pre-processing

This phase starts with a statistical analysis of the data and visualisation of data to understand its key attributes and any significant errors or missing attributes. Table 4.2 shows our Case study contains 500 valid cases and no missing fields.

Table 4. 2 Case processing summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
TransactionValue	500	100.0%	0	0.0%	500	100.0%

Given the nature of the data set, it was applicable to investigate any anomalies in the data using SPSS Boxplot, which provides a quick visual summary of any number of groups, and some evidence regarding the shape of the distribution, the Explore procedure of SPSS offers many options allowing a more detailed look at how groups may differ from each other or from expectation.

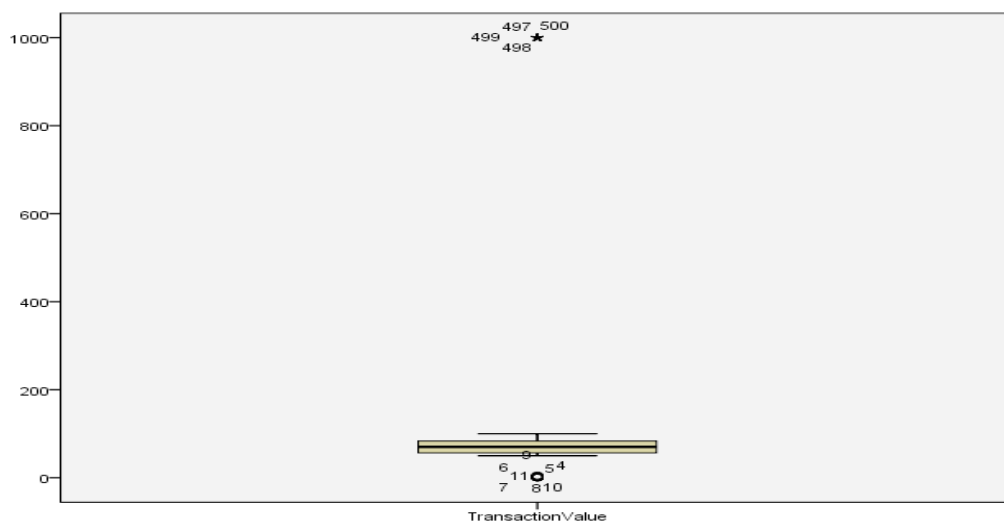


Figure 4. 2 Anomalies in proof of concept data

In Figure 4.2 anomalies are denoted as outliers (O) by SPSS; however, in our study these are described as **point anomalies**, as they refer to the values of transactions such as £1, £2, £3, £5 or £1000. The very low and high amounts spent compared to the normal range, are marked

with an asterisk (*). On the box plot shown here, they are identified by the different markers representing "out" values with a small circle and "extreme values" are marked with a star. This is based on numerical criteria as SPSS uses a step of $1.5 \times \text{IQR}$ (Interquartile range) to define outliers.

4.2.3 Stage 3: Data transformation/coding

Also known as data consolidation, this is a phase in which the selected data is transformed into forms appropriate for mining. For example, the categorical values for the purchase field are grouped and denoted by a set of numerical values (1,2,3,4,5) to represent credit card purchase, cash purchase, debit card purchase, gift voucher purchase and cheque purchase respectively.

Similarly the locations of the supermarkets and names of staff trainers are numerically coded as follows: Birmingham=1, Hull=2, Stafford=3, Oxford=4, Leeds=5, London=6, Manchester=7 Evans=1, Jones=2, Smith=3, Adam=4, Green=5.

4.2.4 Stage 4: Data exploration

Data exploration is concerned with the distribution of the data, and is used to describe the characteristics of variables in sales dataset. Here univariate and bivariate analyses are considered as follows.

i. **Univariate Analysis** explores variables (attributes) one by one. Variables could be either numerical or categorical. There are different statistical and visualisation techniques of investigation for each type of variables. The descriptive statistics for each variable are placed into one table. The tables show a summary of variables with imputed values. The types of statistics shown depend on whether the variable is scale or categorical. Statistics for scale variables include the count, standard deviation, mean, maximum and minimum, of each set of the imputed values. For categorical variables, statistics include count and percent by category for the imputed values.

Table 4.3 gives details of the total number of cases related to the 500 transactions, and the descriptive statistics of the maximum and minimum values.

Table 4. 3 Shows descriptive statics

	N	Minimum	Maximum
TransactionValue	500	1	1000
Valid N (listwise)	500		

Table 4.4 shows 479 transactions classified as non anomalous values and only 21 values are classified as **anomalies**; these 21 anomalies are consider as point anomalies.

Table 4. 4 Shows the frequency statics

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid Nonanomaly	479	95.8	95.8	95.8
Anomaly	21	4.2	4.2	100.0
Total	500	100.0	100.0	

The values in Table 4.5 are identified as 11 low cases of point anomalies (£1, £2, £3, £5) and 10 as high cases of point anomalies (£1000).

Table 4. 5 Shows the frequency of anomalies statistics

High/Low Transaction Values

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid Low Value	11	2.2	52.4	52.4
High Value	10	2.0	47.6	100.0
Total	21	4.2	100.0	
Missing System	479	95.8		
Total	500	100.0		

In Figure 4.3 the bar chart represents the frequency of all transaction values. The mean of the 479 nonanomalies transactions is 70.96 and the standard deviation is 14.917. The mean of the 21 anomalous transactions is 477.1 and the standard deviation is 510.8.



Figure 4. 3 Bar chart of transaction values

ii. **Bivariate Analysis** is the simultaneous analysis of two values (anomalies, nonanomalies) (attributes). It explores the concept of relationship between two variables, their existence and strength, or their differences significance. Table 4.6 shows 479 nonanomalous purchases and 21 anomalous purchases. Cash purchase has the highest number of anomalies, and gift vouchers has one single anomaly.

Table 4. 6 Shows the frequency of non-anomalies/anomalies in the purchase category

Purchase category	Number of		Total
	Nonanomalies	Anomalies	
Credit card (1)	108	5	113
Debit card (2)	96	3	99
Cash (3)	128	7	135
Cheque (4)	80	5	85
Gift voucher (5)	67	1	68
Total	479	21	500

Analysis of Staff ID reveals the greatest value of the transaction values is Staff ID 6 with 3 cases classed as anomalies.

Table 4. 7 Shows the frequency of anomalies/anomalies in staff ID

Staff ID		Number of		Total
		Nonanomalies	Anomalies	
	1	29	2	31
	2	30	1	31
	3	29	2	31
	4	29	2	31
	5	30	0	30
	6	28	3	31
	7	28	2	30
	8	32	0	32
	9	32	2	34
	10	35	1	36
	11	34	2	36
	12	36	2	38
	13	38	1	39
	14	37	0	37
	15	32	1	33
Total		479	21	500

Table 4.8 and Figure 4.4 show that Staff trainer Smith has a total number of 12 cases, associated with anomalous values while the other staff trainers have anomalous transactions ranging between 2 and 3.

Table 4. 8 Shows the number of non anomalies/ anomalies in staff trainer

Sales datasets		Number of		Total
		Nonanomalies	Anomalies	
Staff Trainer	Evans	131	2	133
	Jones	160	3	163
	Smith	0	12	12
	Adam	1	2	3
	Green	187	2	189
Total		479	21	500

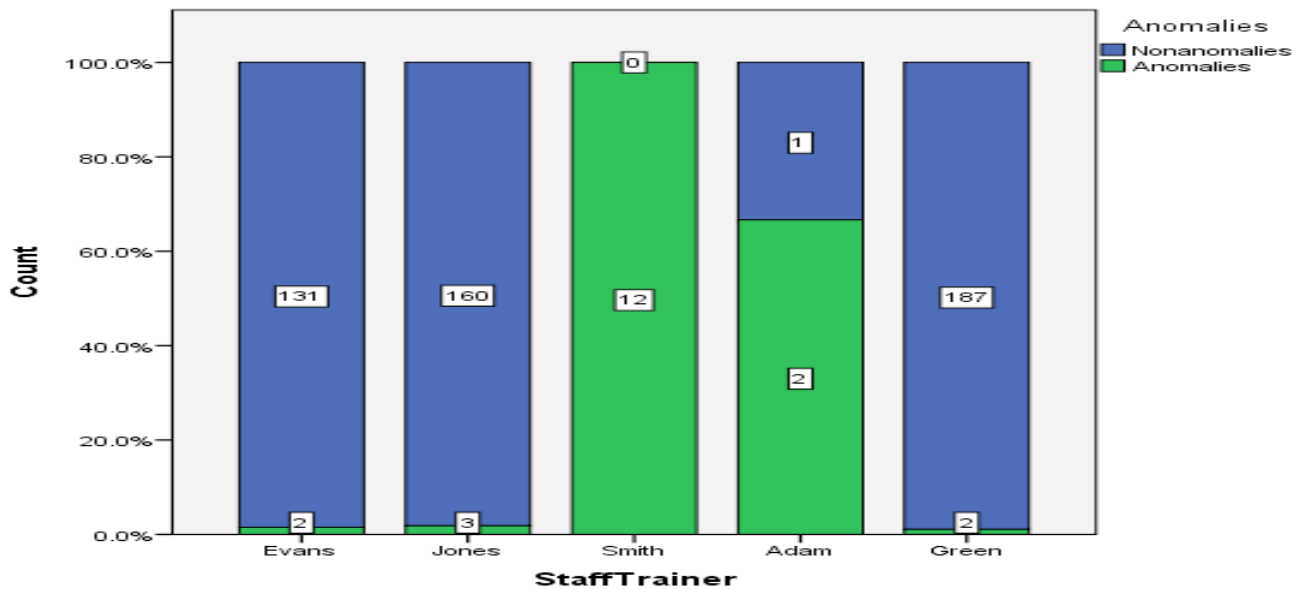


Figure 4. 4 Staff trainer data

4.2.5 Stage 5: Data modeling

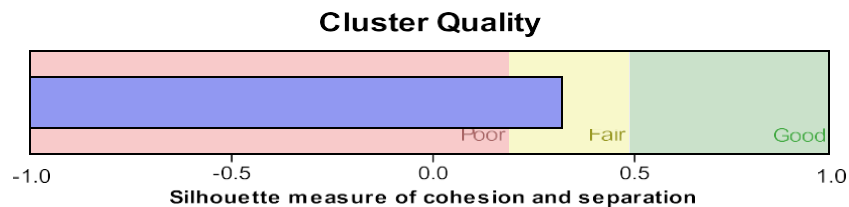
This phase focuses on identifying the appropriate modelling method to be used to capture anomalies. Many data mining algorithms find anomalies to be a side-product of clustering algorithms as clustering aims to partition a set of *data objects* into a number of clusters. Objects with similar features should be grouped together and objects with different features should be placed in divided groups (Fränti & Kivijärvi, 2000).

- Two step cluster analysis

The clustering procedure is based on the SPSS ‘TwoStep Cluster Analysis’. It is a useful for identifying the natural groupings of cases or variables, and it works well with categorical and continuous variables and with very large data files. Table 4.9 shows a summary of the cluster model, including a silhouette measure of cluster cohesion and separation that is shaded to indicate poor, fair, or good results. The results of fair, poor and good are built on the work of Kaufman and Rousseeuw (1990) regarding the interpretation of cluster structures. In the model summary view, a good result equates to data that reflects Kaufman and Rousseeuw's rating as either reasonable or strong evidence of cluster structure. Poor, reflects their rating of no significant evidence and fair, reflects their rating of weak evidence.

Table 4. 9 Model summary

Algorithm	TwoStep
Inputs	5
Clusters	4



The silhouette measures averages over all records, $(B-A) / \max(A,B)$, where A is the record's distance to its cluster centre and B is the record's distance to the nearest cluster centre that it does not belong to. A silhouette coefficient of 1 would mean that all cases are located directly on their cluster centers. A value of -1 would mean all cases are located on the cluster centre of another cluster. A value of 0 means, on average, cases are equidistant between their own cluster centre and the nearest other cluster.

The two step clustering identifies 4 clusters and 5 inputs or predictions representing purchase category, locations, staff ID and staff trainers. As shown in Figure 4.5, 31.8 % of the records are assigned to the first cluster, 38.4 % to the second, 25.6 % to the third cluster, and the fourth 4.2%.

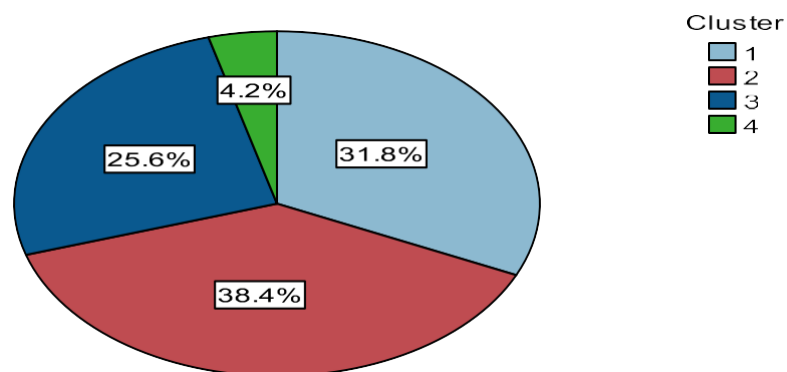


Figure 4. 5 Cluster sizes

The Predictor importance view in Figure 4.6 shows the relative importance of each field in estimating the model. The most important feature is staff trainer and the least important is staff ID.

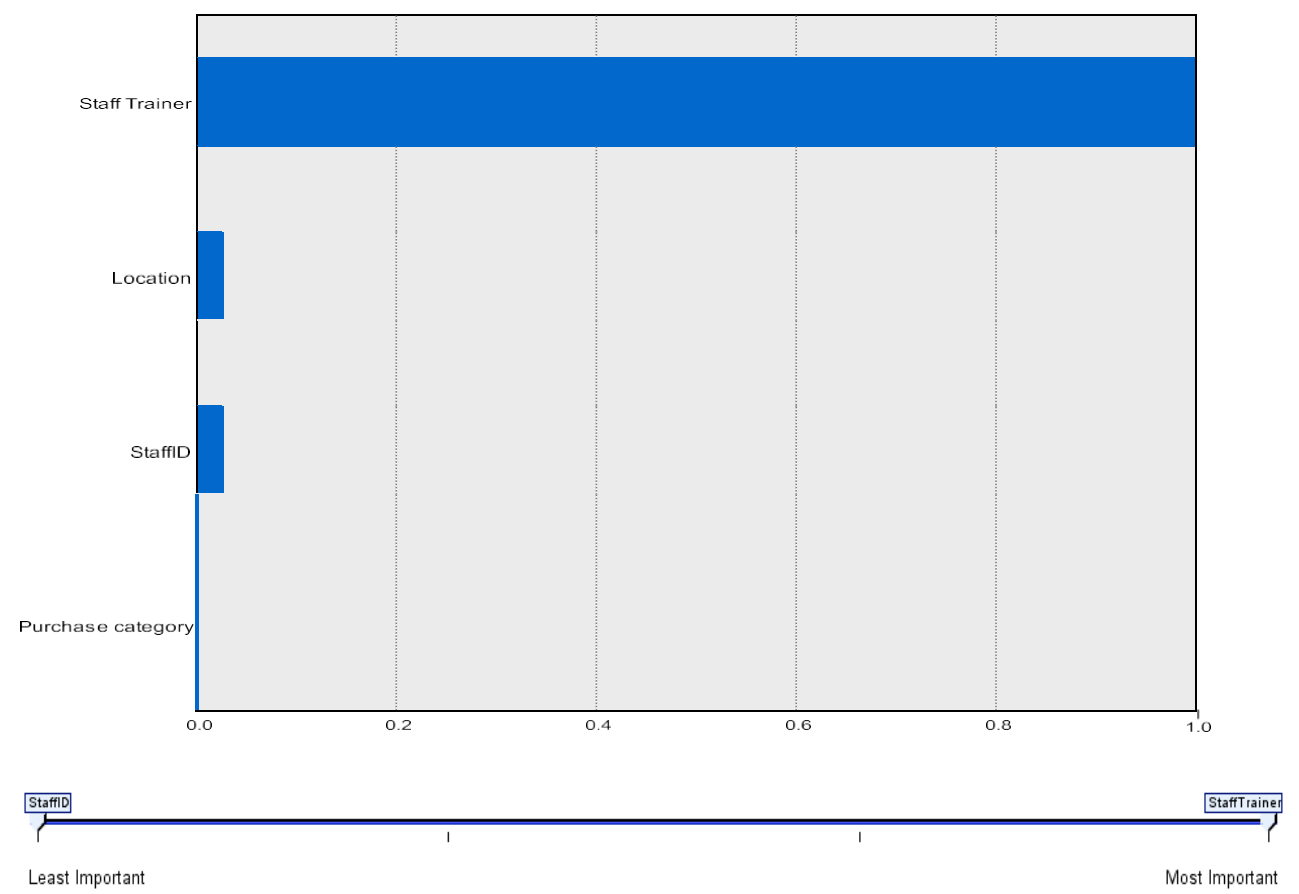


Figure 4. 6 Important features

A guide above in the Figure 4.6 indicates the importance attached to each feature cell colour.

Table 4.10 reveals a cluster-by-features grid that includes cluster names, sizes, and profiles for each cluster. The columns in the grid contain the following information: Cluster, Label, Description, Size and Features. Overall feature importance is indicated by the colour of the cell background shading; the most important feature is darkest; the least important feature is unshaded.

Table 4. 10 The cluster view

Clusters

Input (Predictor) Importance
 1.0
 0.8
 0.6
 0.4
 0.2
 0.0

Cluster	2	1	3	4
Label				
Description				
Size	<div><div style="width: 38.4%;"></div></div> 38.4% (192)	<div><div style="width: 31.8%;"></div></div> 31.8% (159)	<div><div style="width: 25.6%;"></div></div> 25.6% (128)	<div><div style="width: 4.2%;"></div></div> 4.2% (21)
Inputs	StaffID 7 (10.9%)	StaffID 2 (10.7%)	StaffID 10 (13.3%)	StaffID 6 (14.3%)
	Location Leeds (25.5%)	Location Hull (25.8%)	Location Hull (28.1%)	Location Manchester (38.1%)
	Staff Trainer Green (97.4%)	Staff Trainer Jones (100.0%)	Staff Trainer Evans (99.2%)	Staff Trainer Smith (57.1%)
	Purchase category A (27.6%)	Purchase category C1 (27.7%)	Purchase category C1 (27.3%)	Purchase category C1 (33.3%)
	Nonanomales/Anom alies	Nonanomales/Anom alies	Nonanomales/Anom alies	Nonanomales/Anom alies

In the main view from Table 4.10, clusters are sorted from left to right by cluster size, so they are currently ordered 2, 1, 3 and 4 where cluster 4 contains anomalies. This highlights that the staff trainer in cluster 4 is Smith who has 57.1% of anomalies in transactions where purchase category cash purchase C1 is 33.3 % of the total number of anomalies, in location Manchester and the Staff ID is Smith.

The cluster comparison view helps better understand the factors that make up the clusters; it also enables differences to be seen between clusters not only as compared with the overall data, but with each other. Clusters are shown in Figure 4.7 the order in which they were selected; fields are always sorted by overall significance. The background plots display the overall distributions of each feature:

- Categorical features are displayed as dot plots, where the size of the dot indicates the most frequent/modal category for each cluster (by feature).

- Continuous features are showed as boxplots, which display overall medians and the interquartile ranges. Overlaid on these background views are boxplots for selected clusters.
- Square point markers and horizontal lines indicate the median and interquartile range for each cluster. Each cluster is represented by a different colour, shown at the top of the view.



Figure 4. 7 Cluster comparison

Table 4. 11 List of anomalies in the proof of concept data

SaleID	Date	TimeSlot	ioneryPro	chaseCate	Location	nsactionVa	StaffID	Month	rainingLo	taffTraine	HighLow	Anomalies
A059	01-Oct-12	PM	Pencil	2.00	7.00	1.00	11	April	London	3.00	1.00	1.00
A060	01-Oct-12	EVE	encil Rubb	1.00	7.00	1.00	9	April	London	3.00	1.00	1.00
A222	01-Oct-12	AM	Stapler Pin	2.00	7.00	1.00	7	July	London	3.00	1.00	1.00
A246	01-Oct-12	EVE	Pencil	2.00	2.00	1.00	10	Aug	London	5.00	1.00	1.00
A286	01-Oct-12	EVE	pler Remo	5.00	7.00	1.00	13	Aug	London	3.00	1.00	1.00
A397	01-Oct-12	PM	Correction	1.00	2.00	1.00	7	Sept	London	5.00	1.00	1.00
A412	01-Oct-12	AM	Pencil	3.00	7.00	1.00	3	Nov	London	3.00	1.00	1.00
A460	01-Oct-12	AM	Correction	1.00	2.00	2.00	11	Sept	London	1.00	1.00	1.00
A488	01-Oct-12	AM	Stapler Pin	3.00	5.00	2.00	12	Aug	London	2.00	1.00	1.00
A482	01-Oct-12	PM	pler Remo	3.00	2.00	3.00	6	Dec	London	2.00	1.00	1.00
A204	01-Oct-12	AM	ift pen Se	1.00	5.00	5.00	3	July	London	2.00	1.00	1.00
A001	01-Oct-12	Am	aser printe	3.00	7.00	1000.00	1	Jan	London	4.00	2.00	1.00
A028	01-Oct-12	AM	aser printe	4.00	1.00	1000.00	4	Feb	London	3.00	2.00	1.00
A048	01-Oct-12	PM	aser printe	4.00	1.00	1000.00	6	March	London	4.00	2.00	1.00
A105	01-Oct-12	PM	aser printe	3.00	1.00	1000.00	15	May	London	3.00	2.00	1.00
A189	01-Oct-12	AM	aser printe	3.00	1.00	1000.00	12	June	London	3.00	2.00	1.00
A284	01-Oct-12	PM	aser printe	1.00	3.00	1000.00	9	Aug	London	1.00	2.00	1.00
A410	01-Oct-12	EVE	aser printe	4.00	7.00	1000.00	2	Sept	London	3.00	2.00	1.00
A450	01-Oct-12	PM	aser printe	3.00	7.00	1000.00	1	Dec	London	3.00	2.00	1.00

Anomalies are identified in three ways. First, observations that have low probability of being a member of a cluster (i.e. are distant from other cluster members) are identified as anomalies. The probability of 192 records (38.4%) is used as a cut-off point. Second, clusters with small populations of 21 records (4.2%) are considered anomalies, and the third is using SPSS (Boxplot). According to cluster based anomaly detection techniques chapter 2 section (2.6.2) the third assumption, the *Normal data instance belongs to large and dense clusters, while anomalies belong either too small or too sparse clusters*. Techniques based on the above assumption declare instances belonging to cluster as anomalous if size/density is below a threshold.

The **Cell Distribution** shows an increased, more detailed, plot of the distribution of the data in cluster 4; for any feature cell selected in the Clusters main panel (see Figure 4.8). The solid red colour display shows the cluster distribution, while the lighter display represents the overall data.

Note: this display is for cluster 4 only as shown in Figure 4.8:

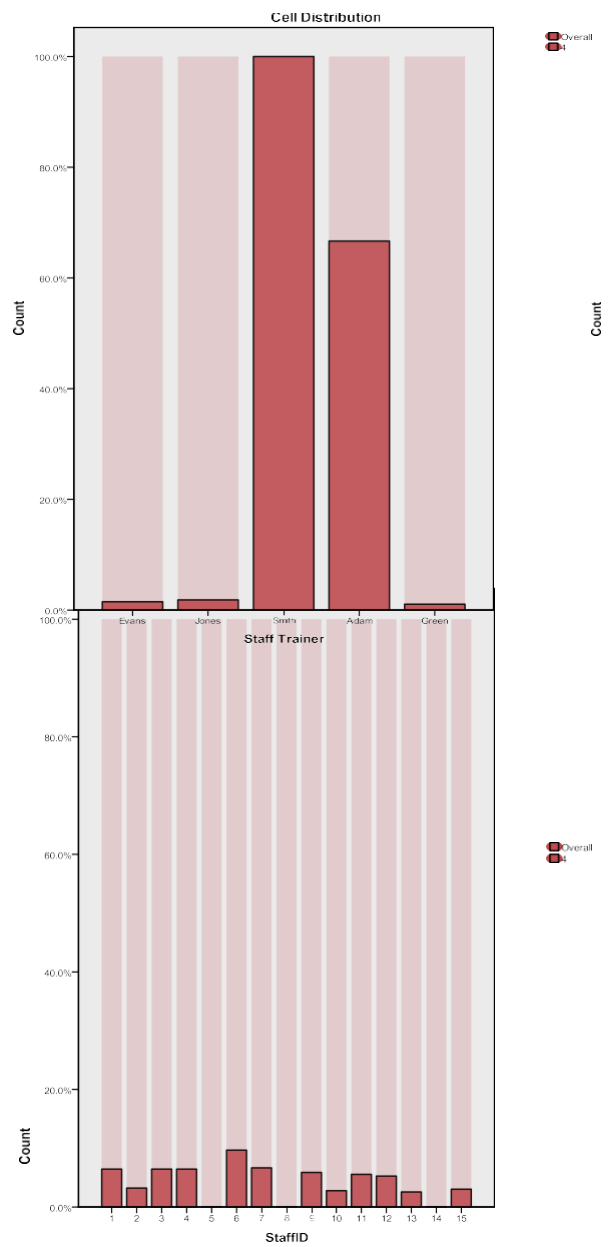
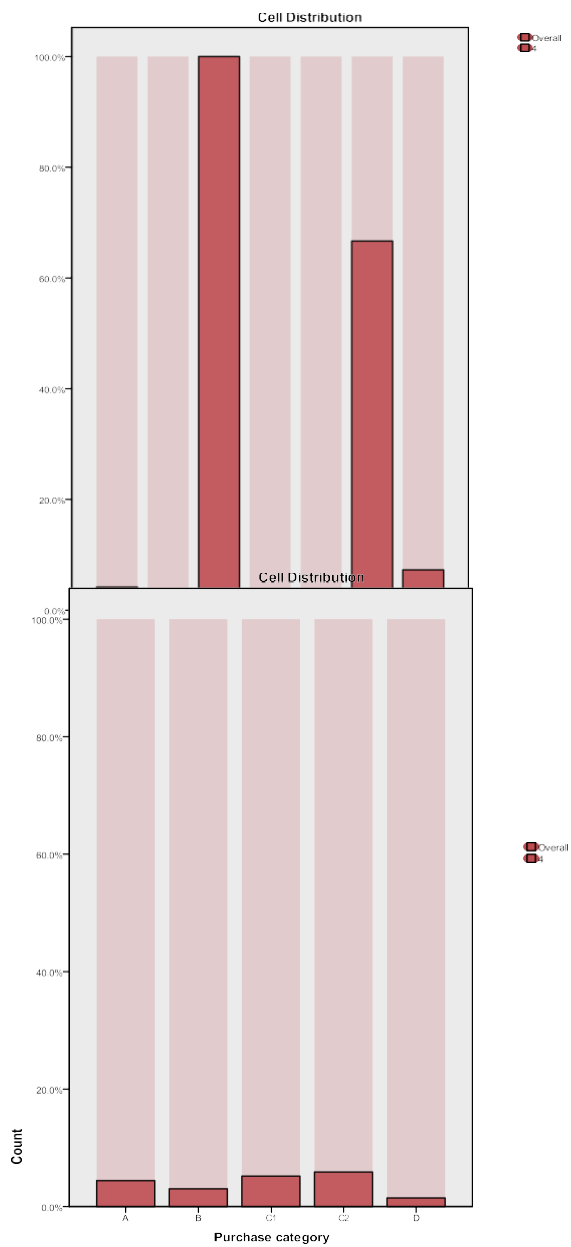


Figure 4. 8 Comparison features in cluster 4

4.2.6 Stage 6: Data evaluation

The requirements for evaluating cluster results are well known in the research community and a number of efforts have been made especially in the area of mining. In general terms, there are three approaches to investigating cluster validity (Theodoridis and Koutroubas 1999). The first is based on external criteria. This implies that we evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a data set and which reflects our intuition about the clustering structure of the data set. The second approach is based on internal criteria. We may evaluate the results of a clustering algorithm in terms of quantities that involve the vectors of the data set themselves (e.g. proximity matrix). The third approach of clustering validity is based on relative criteria. The idea is the evaluation of a clustering structure by comparing it with other clustering schemes, resulting in the same algorithm but with different parameter values.

There are four external criteria of clustering quality (Purity, F measure, Rand index, Mutual information). Purity is a simple and transparent evaluation measure. Mutual information can be information-theoretically interpreted. The Rand index penalises both false positive and false negative decisions during clustering. The F measure in addition supports differential weighting of these two types of errors (Christopher et al., 2008).

In this approach mutual information is used as it involves choosing the clustering that shares the most information with all the other clusterings, such as in Strehl and Ghosh (2002). A measure is therefore needed to quantify the amount of information shared between clusterings. Therefore, the information-theoretic measures form another fundamental class. Such measures work because of their strong mathematical foundation, and their ability to detect non-linear similarities. Based on information theory, mutual information provides a general measure of dependencies between variables. The mutual information used in this research therefore provides a better and more general criterion to investigate relationships between variables.

Mutual information is a quantitative measurement of how much one random variable (B) tells about another random variable (A). In this case, information is thought of as a reduction in the uncertainty of a variable; high mutual information indicates a large reduction in uncertainty whereas low mutual information indicates a small reduction and zero mutual information between two random variables means that the variables are independent. Several parameters must be selected in order to properly run within any given process. The

relationship between variables is integral to correctly determine the working values for the system. If A and B were identical, then all the information derived from obtaining variable A would supply the knowledge needed to get variable B. If two or more variables provide the same information or have similar effects on one outcome, this can be taken into consideration while constructing a model.

The data exploration identifies the clusters and the anomalies; however mutual information gives context to the anomalies and extracts more information. The measure of mutual information between two variables takes all association patterns into account when estimating the extent to which the two variables co-vary with each other. Therefore, this mutual information-based measure probably is a more general way of inferring links in data. MI is used to understand/explain anomalies. This information cannot be obtained by human visualisation especially when the size of data is large. Mutual information is applied to transaction values and staff ID, the result of which was 0.15, for transaction values and staff trainers the result was 0.99. It is also applied to staff ID and staff trainers, which gave a result of 0.32. The mutual information applied to transaction values, staff trainers and staff ID is 0.53.

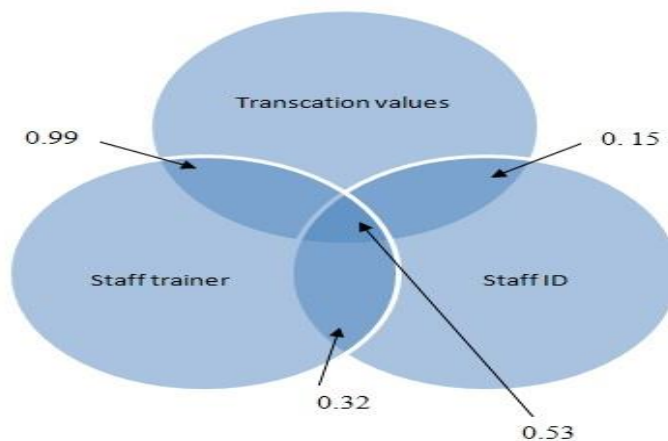


Figure 4. 9 Mutual information of transaction values, staff trainer and staff ID

4.3 Discussion

Case study 1 was designed as a proof of concept investigation to determine whether the approach used had validity. The data set was small and well structured and contained known

anomalies for transaction price. In Case study 1, we operated in an unsupervised setting, where only the normal behaviour is characterised, we used two-step clustering to cluster the data. The data set was designed with intention that the only common factor between anomalies would be by staff trainer/ the transaction values. In applying our method to the data set, the aim was to determine if mutual information could be used to explain the anomalies and the shared link. To identify mutual information Case study 1 we measured the strength of the relationship between elements.

Mutual information has contributed to our understanding of the anomalous features and helped identify links with anomalous behaviour. Data attributes (features) in anomalies detection are divided into two distinct groups: context (or condition) attributes B, and target attributes A. Anomalies detection attempts to identify anomalies in target attributes A with respect to context attributes B. The contextual feature allows identification of patterns that are typical in one context but anomalous in the other. This has led to domain-specific efforts in this area based on factors such as the nature of the data, the type of anomaly, the availability of data labels, and other constraints. In case study1, the transaction value of the transaction dataset is the obvious and straightforward source of anomalies, as it contains very high and very low transaction values compared to the normal range of expenditure for that individual. This approach supported identification of the anomaly but did not explain it. A semantic explanation was needed to understand the anomalies.

Our approach is applied to a small dataset where transactional data is structured data and data patterns are stable. Context plays an important role in anomalies detection, because patterns used to detect anomalies cannot take into account all environmental factors. It is necessary to put each anomaly, once detected, in context. This information can be used to justify the behaviour of an entity. The approach was able to identify a strong anomaly relationship between the transaction values and staff trainers with a mutual information value of 0.99, in particular, other anomalies relationships were identified Table (4.12). For example there are two secondary relationships between the staff ID and staff trainers ($MI=0.53$) and transaction value and location ($MI=0.41$). The weakest relationships are transaction values and purchase category, and transaction value and staff ID.

Table 4.12. Measure of mutual information between two variables.

	Variables	Mutual information (MI)
1	Transaction values & Staff trainers	0.99
2	Staff ID & Staff trainers	0.53
3	Transaction values & Location	0.41
4	Transaction values & Purchase category	0.23
5	Transaction values & Staff ID	0.15

This had not been expected, given that the data set had been constructed around the price/staff trainer anomaly; this highlights the importance of understanding context in data analysis. Case study 1 demonstrated that mutual information could be used to identify and explain anomalies; these anomalies are referred to as point anomalies. The limitations of Case study 1 were that the clustering algorithm used with the data meant that it was difficult to validate the semantic validity of the clusters. As the investigation was carried out with known anomalies in a tightly constrained data set, there was a risk of bias. The results of using the algorithm are appropriate with regard to the concept proof of data, the constructed anomalies values and the small number of anomalies. Based on these observations, the CRISP based methodology can be used to support the semantic investigation of anomalies. It was necessary to demonstrate that the approach could be scaled to real world data volumes and used with inconsistent and/or noisy data. These issues were addressed in Case study 2. Which uses the same approach with a different real world data set.

4.4 Summary

The first sample size of 50 records was not enough to test the anomalies. We then used a larger data sample of about 500000 records from GOWALLA database; however as the anomalies were not known, it became difficult to test the validity of our proposed approach. It was decided to use a set of fictional data referred to as Case study 1, with known anomalies. It consists of 500 records, which included 21 anomalies. This Case study was a useful vehicle to investigate whether mutual information can help identify the anomalies and its source.

Mutual information has contributed to our understanding of anomalous features and helped identify links with anomalous behaviour. The experimentation carried out on Case study 1

advocates the use of mutual information-based measures to represent link strength or proximity between individual anomalies. Case study 1 showed the validity of our approach for dataset where transactional data is structured data and data patterns are stable. The challenge in Case study 2, which is introduced in the next chapter, is to build on the knowledge gained from Case study1 and apply it to the analysis of citation data where the data volumes are much larger, the data patterns are unknown and may be volatile, and the data may be semi-structured. In other words, a move from a stable, limited dataset to a dataset where the boundaries are not known, hence the need to ascertain whether this approach can scale up in this environment.

5 Anomalies Detection: Case Study 2

This chapter discusses the application of the novel approach to Case study 2 to demonstrate how mutual information can help explore and interpret anomalies detection with a different data set and application area. The key challenge for this technique is to apply the same approach to a different real world data set, making use of a different form of data representation, for example graphs to visualise the dataset and a different clustering approach (hierarchical cluster rather than a two-step clustering method). This chapter focuses on a second Case study using a set of co-citation data.

5.1 Anomaly detection methodology applied to Case study 2

The link mining methodology described in chapter 3 is applied to our Case study 2 and includes the following stages: data description, data pre-processing, data transformation, data exploration, data modelling based on graph mapping, hierarchical cluster and visualisation, and data evaluation.

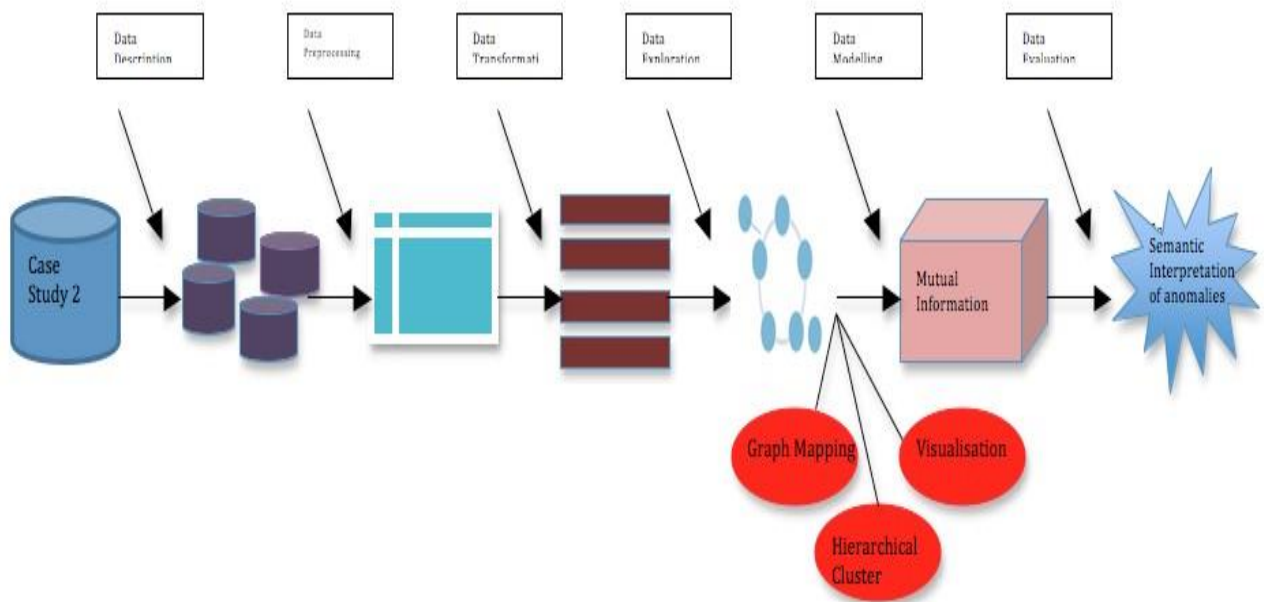


Figure 5. 1 Link mining methodology

This Case study covers the three link mining tasks described in chapter 2. It is an attempt at identifying and clustering objects, representing them into a graph structure and studying the links between these objects.

An important goal in link mining is the task of inferring links that are not yet known in a given network. This Case study aims to use mutual information to interpret the semantics of anomalies identified in our co-citation, dataset which can provide valuable insights in determining the nature of a given link and potentially identifying important future link relationships.

5.1.1 Stage1: Data description

There are several online bibliographic databases where scientific works, documents and their citations are stored. The most important bibliographic databases are the *Web of Science* ISI (WoS), *Scopus*, and *Google Scholar*. 2 extracted 569 records, from *Web of Science*, and stored them in a spreadsheet file. These 569 records include 1001 co-citations from three databases: SCI-EXPANDED, SSCI, A&HCI up to 2011. Each co-citation include the author 'name, journal, cited documents and cited references. The author is the entity that signifies the person who has been involved in the development of the document. An author can be linked to a set of documents, and in a similar way, a document has a group of authors. Also, an author has a linked position in his/her documents. Pairs of citations being cited by a common citing document identified co-citation relationships. The strength of the relationship is based on the number of citing documents that contain the citations. The chance of citations being co-cited increases based on the number of times the citation appears in reference lists of citing documents. Citations contained in a large number of reference lists have a greater chance of being co-cited than citations found in a smaller number of reference lists. Co-citation strength were used to account for the frequencies of citations found in the reference lists of citing documents (see Figure 5.2 &5.3 for examples of the data).

5.1.2 Stage 2: Data pre-processing

The data from the bibliographic sources contain a number of errors, such as misspelling in the author's name, in the journal title, or in the references list. Occasionally, additional information has to be added to the original data, for example, if the author's address is incomplete or wrong. For this reason, the analysis cannot be applied directly to the data

retrieved from the bibliographic sources; a pre-processing task over the retrieved data is required, to improve the quality of the data and the analysis. A set of pre-processing tasks is applied to prepare the data and is described below.

- *Data reduction* aims to select the most important data, which is normally an extensive task. With such a quantity of data, it could be difficult to obtain good and clear results in the relationship. For this reason, it is often conducted using a portion of the data. This portion could be, for example, the most cited articles or the most productive authors. For the journals with the best performance metrics, the most cited reference was adopted for subsequent analysis.

- *Detecting duplicate and misspelled items*: There are items in the data that represent the same object or concept but with different spelling, for example, an author's name can be written in different ways (e.g., Zakia.II; Il Agure Zakea), and yet each spelling represents the same author. In other cases, a concept is represented with different words (lexical forms) or acronyms, and yet refers to the same concept. To improve data quality, first authors' initials, are kept and converted from lower to upper case to maintain consistency. The first author 'name is used in our analysis.

5.1.3 Stage 3: Data transformation

Several relations among the nodes can be established. The focus in Case study 2 was on co-citation in the bibliometric technique taxonomy (see Table 5.1). The most common nodes of analysis are authors, journals, documents, cited references, and key words. Co-occurrence of nodes of analysis are used to investigate the data. The similarity between the nodes of analysis is usually measured counting the times that two nodes appear together in the documents. The nodes of analysis used in Case study 2 are author, citation document and journal.

Table 5.1 Bibliometric techniques taxonomy

Bibliometric technique		Unit of analysis used	Kind of relation
Bibliographic coupling	Author	Author's oeuvres	Common references among author's oeuvres
	Document	Document	Common references among documents
	Journal	Journal's oeuvres	Common references among journal's oeuvres
Co-author	Author	Author's name	Authors' co-occurrence
	Country	Country from affiliation	Countries' co-occurrence
	Institution	Institution from affiliation	Institutions' co-occurrence
Co-citation	Author	Author's reference	Co-cited author
	Document	Reference	Co-cited documents
	Journal	Journal's reference	Co-cited journal
Co-word		Keyword, or term extracted from title, abstract or document's body	Terms' co-occurrence

Table 5.1 shows the taxonomy of the most common bibliometric techniques according to the unit of analysis used and where the established relationship among them is presented. Different aspects of a research field can be analysed depending on the selected nodes for analysis. Additionally, a link can be used to attain the relation among nodes, the extraction of co-citation network by using *BibExcel*, in order to help with citation studies, and bibliographic analysis, in particular:

1. Convert to dialog format/convert from Web of Science.

A bibliographic record consists of a number of fields used to index the actual text, its subjects and descriptive data. As showed above, when working with BibExcel we usually transform the initial data to the dialog format in Figure 5.2 more specifically the format for Science Citation Index. Common data between records are thus structured in univocal metadata fields, such as publication titles in the title field, authors in the author field, and references in the reference filed.

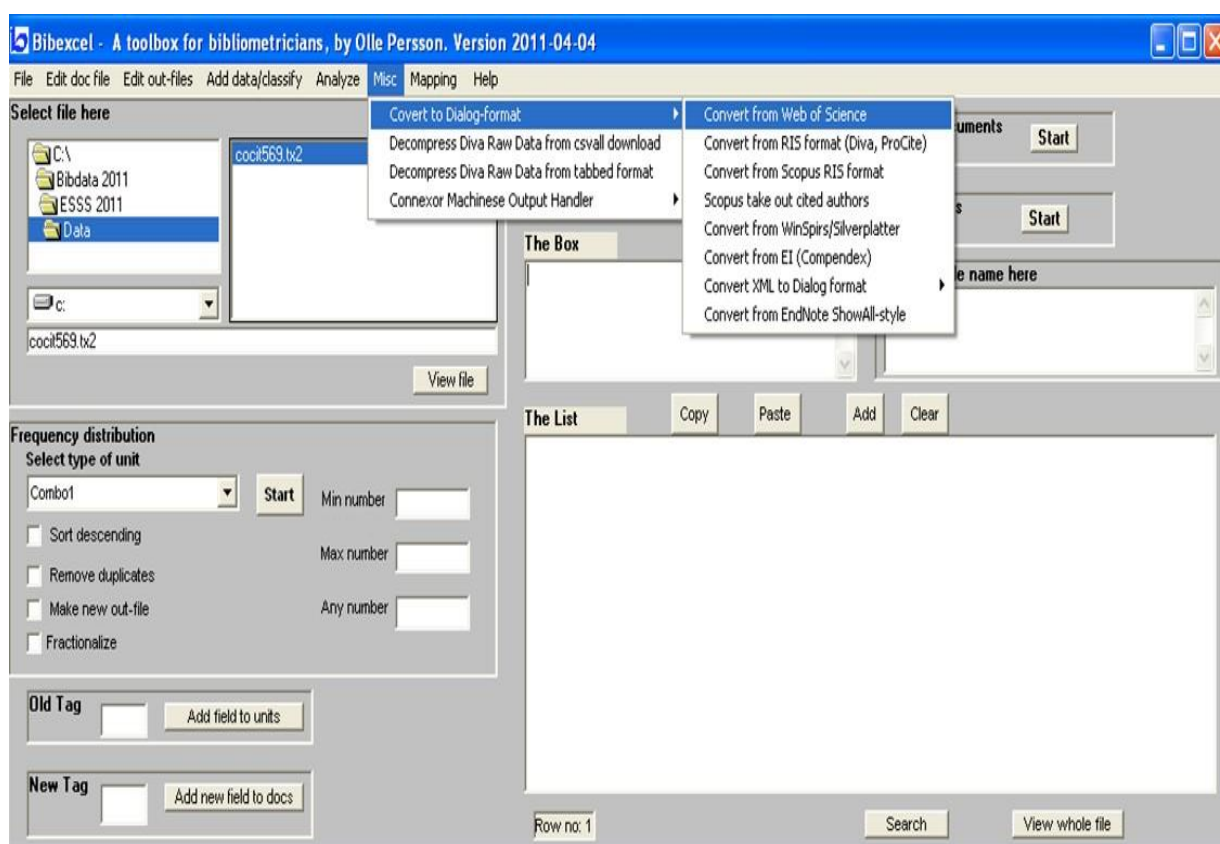


Figure 5. 2 Convert to dialog format

2. Extracting data from CD-field (citation-documents) where the relations of the different entities related with each document (authors, year, vol., page, and journal) are stored. We may want to familiarise ourselves with the structure of the Doc-file. BibExcel keeps track of where the bibliographic record begins and ends by looking for a "|" (double-spike). In addition each record is composed of numerous bibliographic fields and BibExcel keeps track of where the bibliographic fields begin by field tags. Each bibliographic field ends with a "|" (single spike). In fields with multiple units, units are separated from each other with a delimiter. For most bibliographic fields the field delimiter is a semicolon, as shown in Figure 5.3.

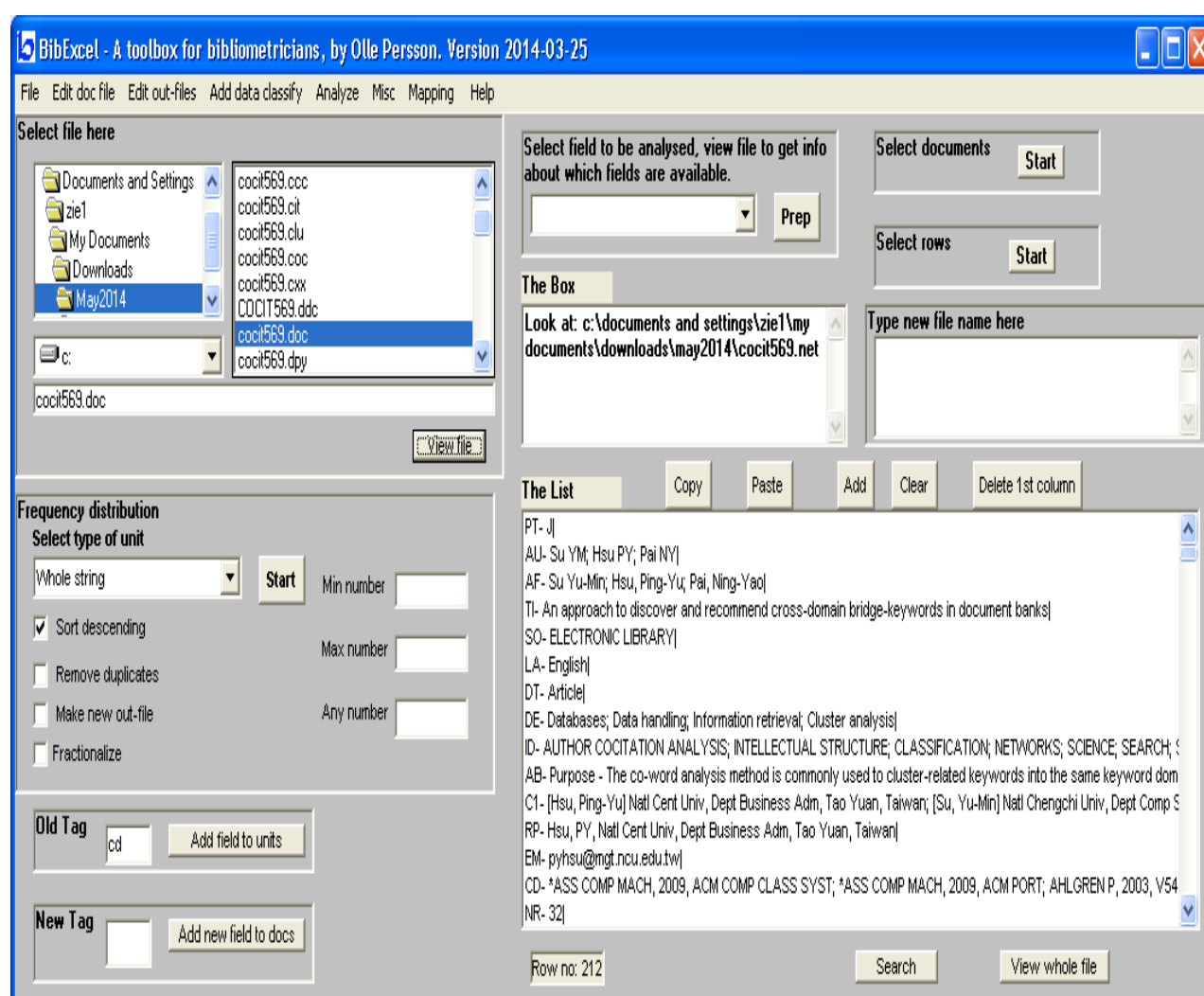


Figure 5. 3 Extracting data from CD-fields (citation-documents)

3. To improve data quality, only the first authors' initials are retained (see Figure 5.4).

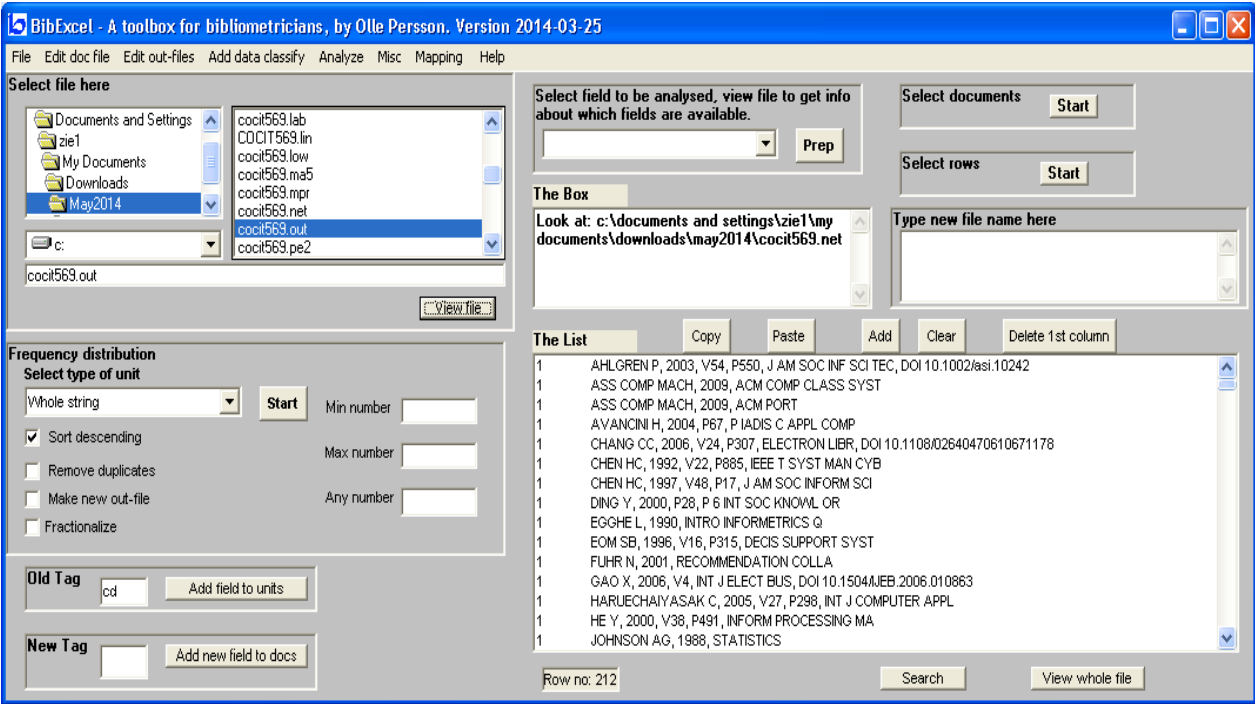


Figure 5. 4 Retaining first authors' initials

4. Convert upper /lower case to improve cited reference strings (see Figure 5.5).

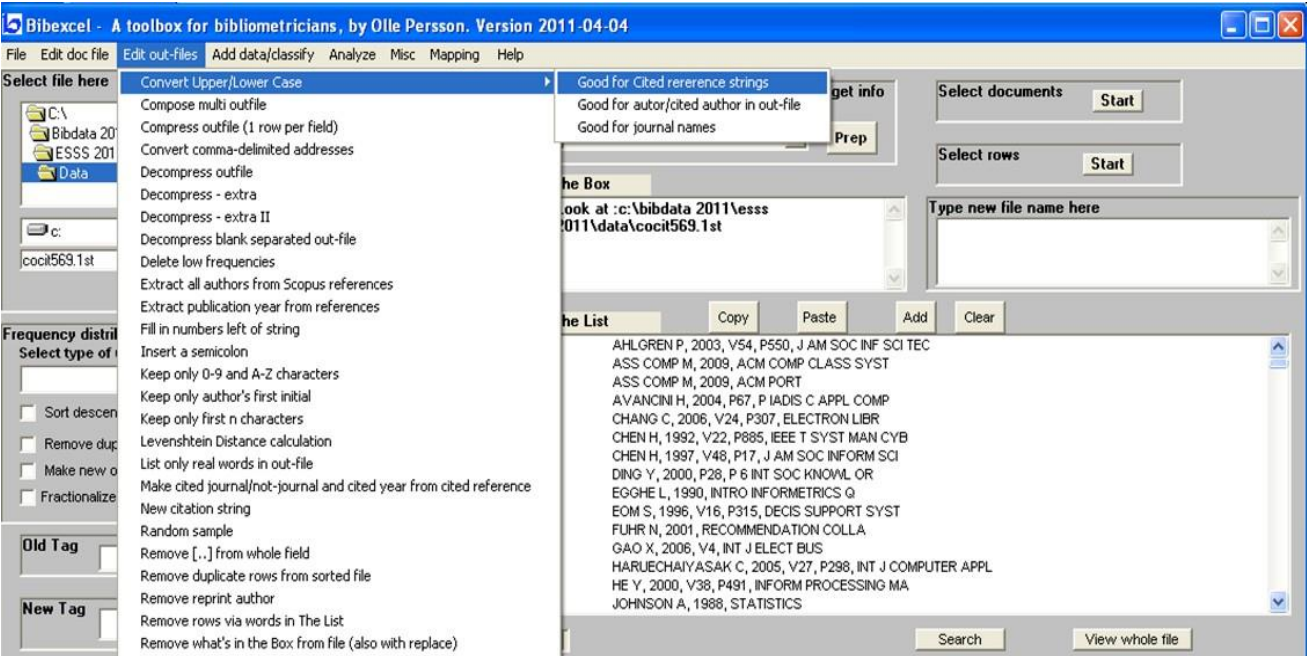


Figure 5. 5 Convert upper/lower cases

5.1.4 Stage 4: Data exploration

Once the network of relationships between the selected nodes has been built, an exploration is applied to the data to derive similarities from the data. For instance, if a co-citation analysis is performed and various clusters are detected, then a label would be set to each one. This label should be selected using the most important document terms of the cluster.

a) Computing frequencies of citations

Making an OUT-file is always the first step when analysing bibliographic data with BibExcel (see Figure 5.6). When making the OUT-file, specific bibliographic fields need to be selected, from which the OUT-file will be constructed. Depending on which bibliographic fields are chosen as a unit when the OUT-file is created, the frequency calculation function in BibExcel offers many different selections. Such as, if the file name: OUT-file consists of a cited document, BibExcel can make a substring search and only count a specified part of the cited document, such as cited author or cited journal.

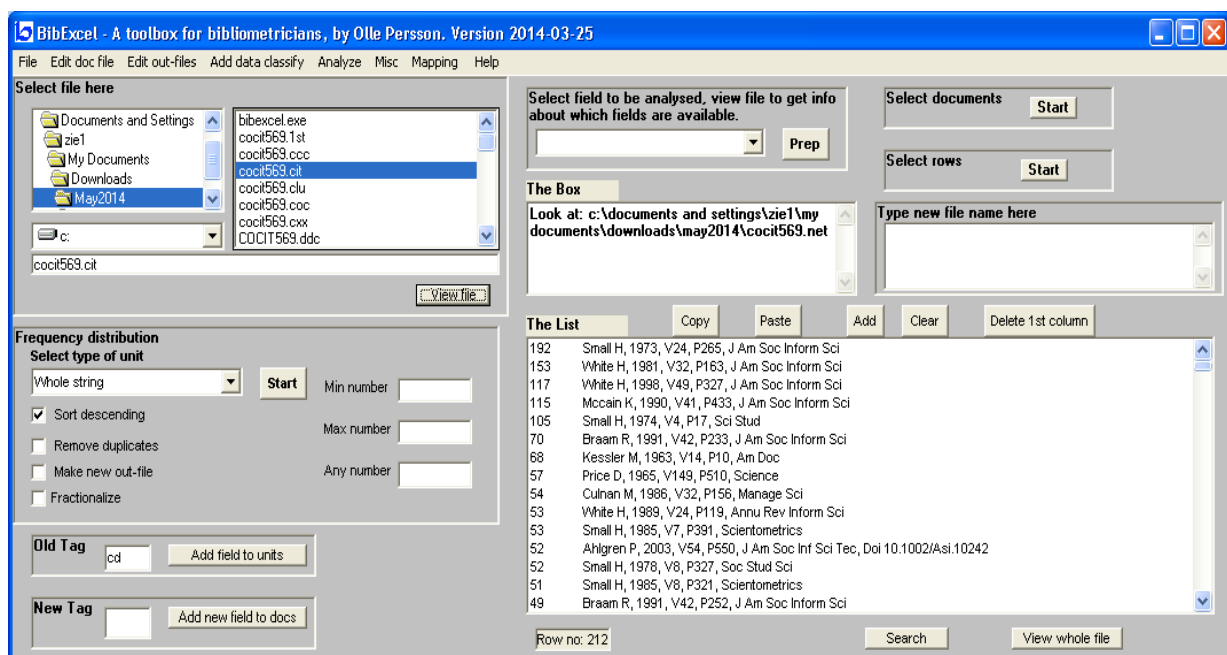


Figure 5. 6 The frequency

b) Making co-citations

Co-citation is a semantic similarity measure for documents that makes use of citation relationships. The definition of co-citation is the frequency with which two

documents *are* cited together by other documents (Small, 1973). If at least one other document cites two documents in common these documents are co-cited. The higher the co-citation strength, the more co-citations two documents receive and more likely they are semantically related (see Figure 5.7).

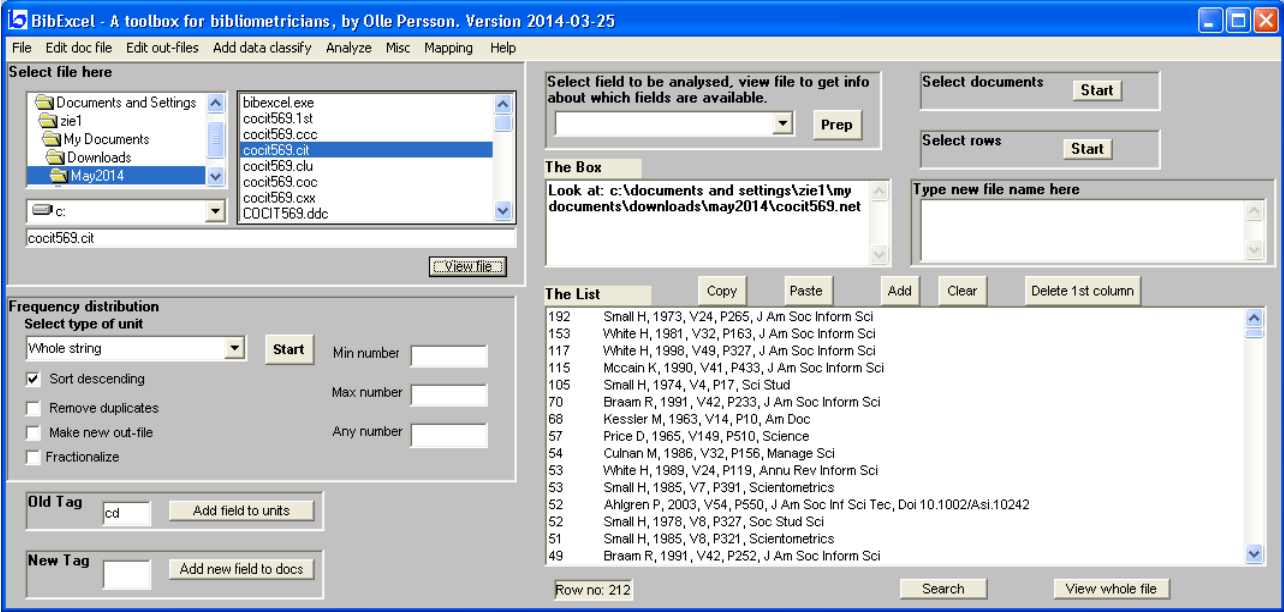


Figure 5. 7 Making co-citations

3. Make co-occurrences *pairs* via the list box (Figure 5.8).

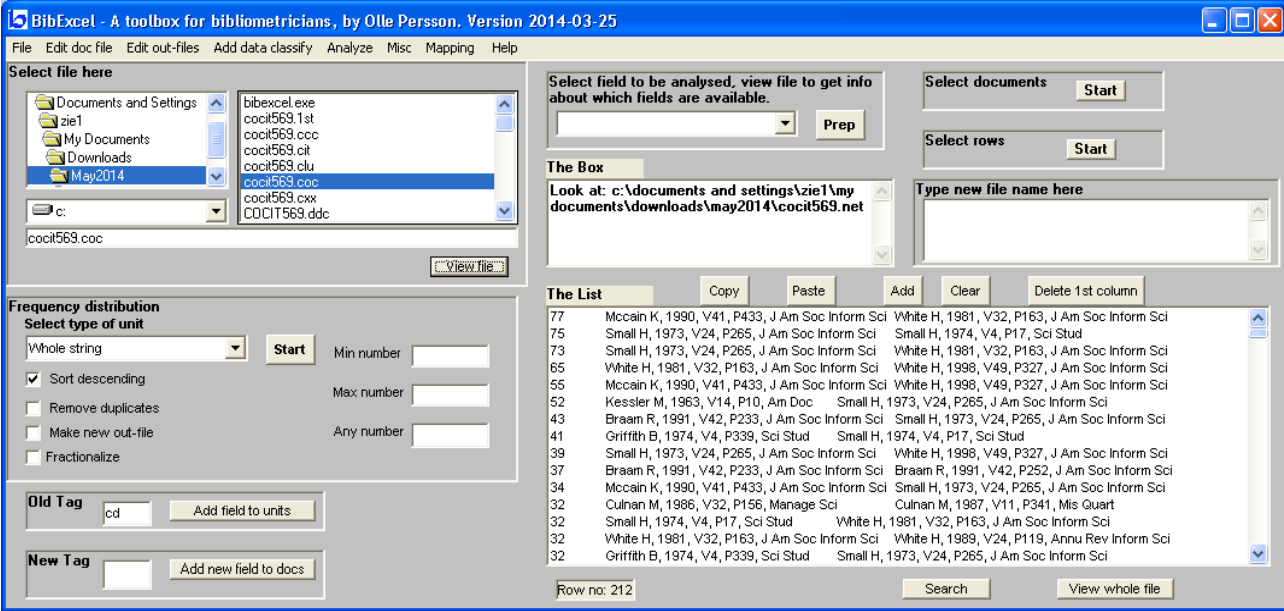


Figure 5. 8 Making co-citations pairs

The menu analysis presented contains a number of specialised functions permitting the analyses of citation networks and, perhaps most importantly, a range of different co-occurrence analyses. We will therefore focus on co-occurrence analysis – how to prepare the data and how to perform co-occurrence analyses.

Co-occurrence analysis is the study of mutual appearances of pairs of units over a consecutive number of bibliographic records. Therefore, the unit of analysis in the OUT-file defines the type of co-occurrence analysis. For example, an OUT-file that lists the individual authors from each record in the Doc-file would be the basis for a co-author analysis. The matching routine used to match pairs of units must therefore be performed on the OUT-file. It is the nodes in the individual documents and their frequency across all documents that must be generated.

Many individual units will have very low frequencies. Such units are often unimportant in co-occurrence analysis, as their mutual relationships will be trivial owing to low frequencies (Olle, 2010). It is therefore a very optimal idea to use individual frequency as an inclusion criterion for the analysis. Furthermore, such a criterion also speeds up the generation of co-occurrence pairs, since this can be a resource demanding routine depending on the number of units to match. As a result, the analysis is focused on those documents which are cited-by at least 10 other authors.

5.1.5 Stage 5: Data modelling

The modelling step is the most important stage. The co-cited data is represented first using a graph representation for visualisation purposes. BibExcel is used to produce *net-files* for co-citations, which are converted for further analysis and visualisation with **VOSviewer** (See Figure 5.9). The VOSviewer tool is used to build a map based on a co-occurrence matrix. (Van Eck and Waltman, 2009a, 2009b). The VOS viewer map created for Case study 2 is given in Appendix B.

5.1.5.1 Graph analysis of co-citation data

Anomalies represent significant deviations from ‘normal’ structural patterns in the underlying graphs. This description is lengthy because much is involved in its preparation, measurement, results and expressing the differences between the groups in some way (the statistic test), and choosing an inference procedure built on that statistic. Each pattern is under the control of the experimenter or observer and each is important. The concept of finding a pattern that is ‘similar’ to frequent, or good pattern is different from most approaches that are looking for unusual or ‘bad’ patterns. There is no universal definition of the problem, as it depends heavily on: The application domain and the properties in addition to the properties of the graph under consideration.

The main goal of anomalies in graphs is to highlight unusual relationships in the graphs by representing them as edges between regions of the graph that rarely occur together. In a citation network, two co-authors who are drawn from groups that usually do not work together may sometimes publish together (cross-disciplinary papers). Such anomalies provide unique insights about the relationships in the underlying network.

Anomalies may be modelled in different ways depending upon the abnormality of either the nodes in terms of their relationships to other nodes, or the edges themselves. In such cases, in Figure 5.10 below a node, which illustrates irregularity in its structure within its region, may be considered as an anomaly (Akoglu et al., 2010). Also, an edge which connects different communities of nodes may be considered a relationship or community anomaly (Aggarwal et al., 2011) and (Gao et al., 2010). Figure 5.10 (a) contains a case of a node anomaly, because node 5 has an unusual locality structure, which is significantly different from the other nodes as (Chen C, 1998, V9, P267, J Visu) in the map. Figure 5.10 (b) Node 5 is that disconnected and is far away from other cluster members as (Zitt M, 1994, V30, P333, Scien)in the map. On the other hand, the edge (2, 4) in Figure 5.10 (c) may be considered a relationship anomaly or community anomaly, because it connects two communities, which are usually not connected to one another as (Kessler M, 19963, V14, P10, Am) in the map. Hence, in the graph data, there is significantly more difficulty and flexibility in terms of how anomalies may be defined or modelled.

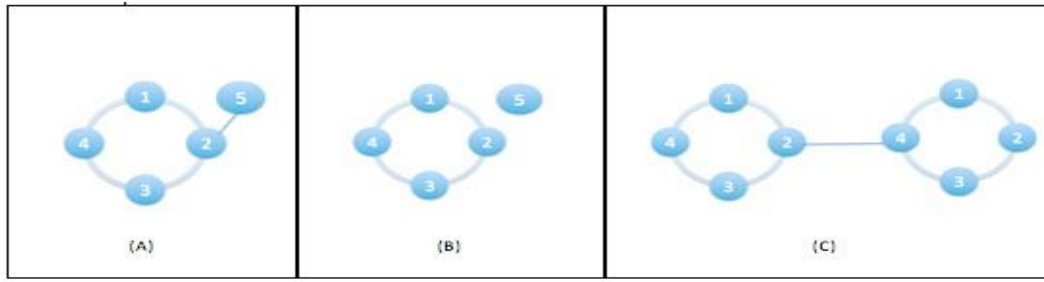


Figure 5. 10 Cases of node anomaly

5.1.5.2 Hierarchical Cluster

A crucial step to evaluate whether mutual information-based measures can be effectively used to represent strength of group ties in network analysis is to examine the extent to which the network structures derived from mutual information-based measures resemble the true network structures. Thus, hierarchical cluster is introduced in the current study for the purpose of network structure inference. Hierarchical cluster is one of the many strategies that have been used to visualise the relationship among elements of a network and to make inferences on the overall structure of the network from proximity data among those elements (Aghagolzadeh et al., 2007; DeJordy, et al., 2007). A hierarchical clustering is a nested sequence of partitions. This method works on both bottom-up and top-down approaches (agglomerative and divisive). In this experiment a bottom-up approach is selected. Hierarchical clustering uses different metrics such as Euclidean distance, Squared Euclidean distance, Manhattan distance and maximum distance, in this experiment the maximum distance metrics was used (Hastie et al., 2009) which measures the distance between two elements and the linkage criteria, which specifies the dissimilarity in the sets as a function of the pair-wise distances of observations in that sets. Given matrix of n elements, the primary goal of hierarchical clustering analysis is to find a partition hierarchy. This analysis is usually performed as beginning from a full partition where each element forms a subgroup; elements are grouped together step by step. At each step, the joining of two subgroups is taken to form a larger group. A new group formation at each step should ensure maximum preservation of relationships between elements as provided in the matrix. The whole partition hierarchy can be created at the n^{th} step and all clusters along with their substructures can then be detected. This experiment applied MATLAB software to the hierarchical clustering for case study 2; a

summary of the algorithm is given below (Day & Edelsbrunner 1984). Further program is given in Appendix C.

```

Given:
A set  $X$  of objects  $\{x_1, \dots, x_n\}$  (1)
A distance function  $dist(c_1, c_2)$  (2)
for  $i = 1$  to  $n$ 
     $c_i = \{x_i\}$ 
end for
 $C = \{c_1, \dots, c_n\}$ 
 $l = n+1$ 
while  $C.size > 1$  do (3)
    -  $(c_{min1}, c_{min2}) = \text{minimum } dist(c_i, c_j)$ 
    - remove  $c_{min1}$  and  $c_{min2}$  from  $C$ 
    - add  $\{c_{min1}, c_{min2}\}$  to  $C$ 
    -  $l = l + 1$ 
end while

```

Given a set of 1001 items to be clustered, and distance (or similarity) matrix, the hierarchical clustering algorithm:

1. Allocates each observation to its own cluster based on author (Co-citation data).
2. Finds the closest (most similar) pair of clusters and merge them into a single cluster, (so there is one less cluster).
3. Computes distances (similarities) between the new cluster and each of the old clusters.
4. Repeats step 2 and step 3 until all items are clustered into a single cluster of size X .

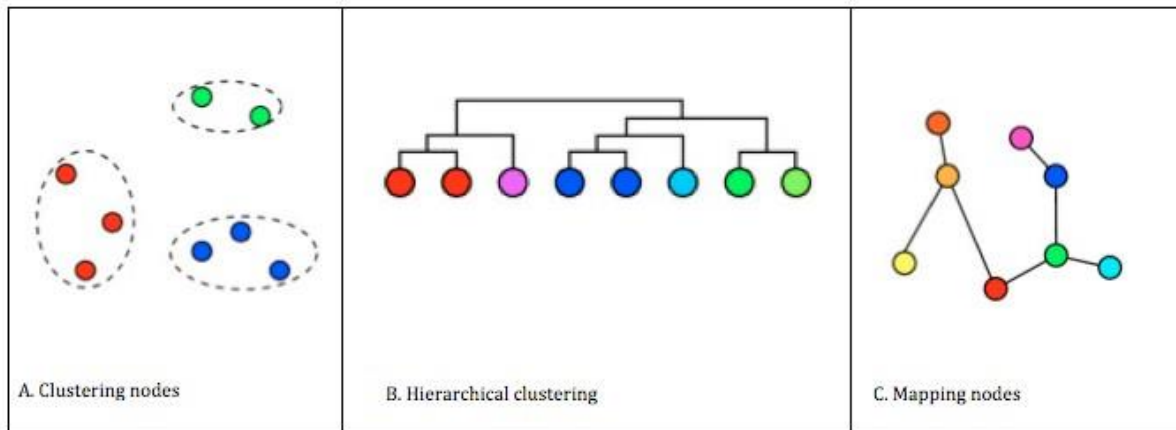


Figure 5. 11 Clustering

A clustering model attempts to determine the region of the network, which is dense in terms of linkage behaviour (see Figure 5.11). In some cases it is also possible to integrate the content behaviour into the detection process. Clustering algorithms was used to group data into 5 different clusters. The clustering grouped 193 nodes, into 5 clusters. The largest cluster is cluster 1 with 58 items and cluster 5 is the smallest with 19 items.

Co-citation is defined as the frequency with which two documents are cited together by other documents. If one other document cites two documents in common these documents are co-cited. The higher co-citations two documents receive, the more their co-citation strength, and are semantically related, which can be related to the results from the mapping nodes. Where cluster 1 shows high co-citation frequency indicating higher co-citation strength, cluster 5 has a low co-citation frequency indicating lower co-citation strength. The relationship strength is based on the number of citations the two citing documents have in common. After the creation of author co-citation pairs, the co-citation link strength (Garfield, 1980) is calculated using the following formula:

$$\text{Link Strength } (AB) = X / (Y - X)$$

Where X is the number of co-citations of author A and author B, Y is the sum of the total number of citations of A and the total number of citations of B. This formula normalises the co-citation link strength by taking into account the total number of citations for both A and B (see Table 5.3). In item 1 (Small H, 1973) the link strength is 1818 indicating that it is present in cluster 1 and is more co cited, however item 193 (Farhoomand A, 1987) is shown to have the lowest link strength of 50 and is present in cluster 5 indicating that it less co cited.

Table 5. 1 Table of link strength

No	Items	Total link strength
1	Small H, 1973, V24, P265, J Am Soc Inform Sci	1818
2	White H, 1981, V32, P163, J Am Soc Inform Sci	1757
4	White H, 1981, V32, P163, J Am Soc Inform Sci	1320
3	Mccain K, 1990, V41, P433, J Am Soc Inform Sci	1319
....
....
193	Farhoomand A, 1987, V18, P48, Data Base	50

5.1.5.3 Visualisation

Analysis of networks has been widely used in a great number of areas to understand relationships between different entities in a network, as well as behaviour of a network as a whole due to the interactions between entities within it. Researchers have conducted observations and developed, experiments on a variety of network analysis techniques including graphical visualisation, statistical inference and computational algorithms, and built a number of mathematical models in an effort to understand and predict the behaviour of a network (Newman, 2003). Figure 5.12 explains how both mutual information and visualisation are used in Case study 2 to validate the approach.

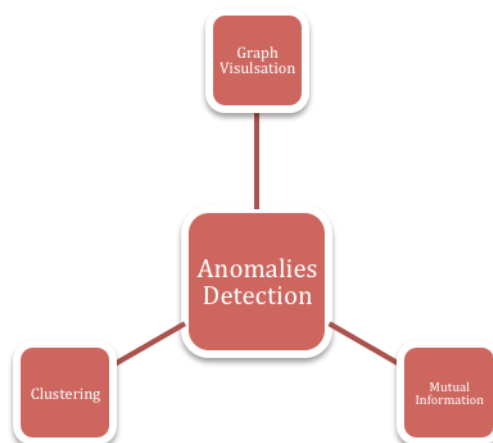


Figure 5. 12 Validating the approach

Co-citation data can be used to study relations among authors or journals; it can be used to construct the maps that provide a visual representation of the structure of a scientific field. Usually, when using co-occurrence data, a transformation is applied first to the data. The aim of such a transformation is to derive similarities from the data. For example, when researchers study relations among authors based on co-citation data, they typically derive similarities from the data and then analyse these similarities using hierarchical clustering.

The visualisation helps provide a clear understanding and better representation of the output map represented at co-citation (see Figure 5.9). The resulting map visualises a set of objects and the relations among the objects. Many different types of visualisations can be used. One difference is between distance-based visualisations and graph-based visualisations. In distance-based visualisations, the distance between two nodes reflects the relation between the nodes. The smaller the distance between two nodes, the stronger the relation between the nodes. On the other hand, in graph-based visualisations in Case study 2, the distance between two nodes does not reflect the relation of the nodes. Instead, drawing lines between nodes from the map typically indicates relations between nodes; the most basic way to visually group nodes is to use colours. If items have been assigned to clusters, the colour of the circle of an item can be determined by the cluster in which the item belongs. Item cluster is calculated and translated into colours using a colour scheme. By default, VOSviewer uses a red-green-blue colour scheme (see Table 5.3). In Case study 2, the relation between nodes is shown by colour and size.

In this colour scheme, red corresponds with the highest item density in cluster 1 and yellow corresponds with the lowest item density in cluster 5. Furthermore the node size denotes the number of received citations (White H, 1981, V 32, P163, JAm) being the largest node in the map, while (Chen C, 2001, V34, P65, Compute) is the smallest node. This can give a great insight into the relations inside a group and between different groups.






5.1.6 Stage 6: Data evaluation

The main objective of visualising the co-citation data using graphs is to highlight unusual relationships in the graphs by representing them as edges between regions of the graph that rarely occur together. In citation network, two co-authors who are drawn from groups that usually do not work together may sometimes publish together (cross-disciplinary papers).

Such anomalies provide unique insights about the relationships in the underlying network. Hawkins (1980) defines an anomaly detection based graph as finding “graph objects (nodes/edges) that are rare and that differ significantly from the majority of in the reference graph nodes.” Graph investigation technique permits the user to filter out nodes based on visual and semantic attributes. The method allows filtering-out nodes by their groups (colours). In addition, the method adopted in this research allows easy modification of filtering options, which may be dependent on other attributes. Each paper in the collection is associated with the authors who wrote it and the references it cites. Cluster 5 consists of papers, which covers *visualisation of literature technique*. All of the element were based on three types of literature, bibliometrics, scientometrics, and informetrics. The mutual information for cluster 5 is 0, which confirms that the elements of that cluster are not linked to other clusters and are considered as **collective anomalies** with respect to the entire dataset. Cluster 1 whose mutual information is 93 confirms that the elements of this cluster share common characteristics/domain areas, which are *Library and information science techniques*.

In Table 5.3 where cluster 1 shows high mutual information indicating higher co-citation strength, cluster 5 has a low mutual information indicating lower co-citation strength.

Table 5. 2 Result of mutual information

	Clusters	Items	Colour	Mutual information
1	Cluster1	58		0.93
2	Cluster2	49		0.82
3	Cluster3	38		0.63
4	Cluster4	29		0.43
5	Cluster5	19		0.00

We applied mutual information to detect anomalies in the context of co-citation, using the equation below:

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x) p(y)} \right)$$

We computed the mutual information MI (X, Y) between two attribute sets X and Y, and only where the mutual information is greater than a threshold. We define X and Y to be dependent on:

$$I(X, Y) \geq \beta\mu$$

Where, $\beta\mu$ is a threshold parameter set to 0.1 in our Case study 2. Thus, for a given node we consider all pairs of dependent and mutually exclusive subsets having up to n nodes, and calculate the corresponding γ -values. A ratio of the form:

$$\gamma = \frac{p(X_t, Y_t)}{p(X_t)p(Y_t)}$$

It has been proposed as a measure of suspicious coincidence by Barlow, (1989). It conditions those two nodes X and Y should be combined into composite nodes XY if the probability of their joint appearance $P(X, Y)$ is much higher than the probability expected in case of statistical independence $P(x) P(Y)$. Here high values of γ are interesting as it signifies a suspicious coincidence of the events co-occurring. From Table 5.3 above we can conclude that cluster 1 has the highest mutual information calculation value 0.93, in comparison to cluster 5 that has the lowest mutual information calculation value 0.0. This indicates that in cluster 1 there has been a strong relationship among the nodes; however, in cluster 5 the relationship among the nodes is weak. We are interested in exactly the opposite situation, where low γ values signifies that the events do not co-occur naturally. If they are observed together, it is then treated as an anomaly. An unusually low value of the ratio suggests a strong negative dependence between the occurrences of nodes in the data. This also ensures we have seen enough cases of nodes to support the theory of negative dependence.

5.2 Discussion

Case study 1 identified a number of issues including the difficulties of confirming the semantic validity of the clusters. If the approach were to be valid when used with a data set where the anomalies and relationships are unknown, it was necessary to demonstrate that the approach could be scaled to real world data volumes and used with inconsistent and/or noisy data and with other clustering algorithms. Case study 2 addresses these issues. Case Study 1 used a two-step clustering algorithm but the clustering approach used in Case study 2 was

hierarchical clustering. Using the bibliographic data, this approach created 5 clusters. Cluster 1 was found to contain data with the strongest links and cluster 5 to contain data with the weakest links. Applying mutual information, we were able to demonstrate that the clusters created by applying the algorithm reflected the semantics of the data. Cluster 5 contained the data with the lowest mutual information calculation value. This demonstrated that mutual information could be used to validate the results of the clustering algorithm.

It was necessary to establish whether the proposed approach would be valid if used with a data set where the anomalies and relationships were unknown. Having clustered and then visualised the data and examined the resulting visualisation graph and the underlying cluster through mutual information, we were able to determine that the results produced were valid, demonstrating that the approach can be used with the real world data set. Analysing each of the clusters, and the relationships between elements in the clusters was time consuming but enabled us to establish that the approach could be scaled to real world data and that it could be used with anomalies which were previously unknown.

We found with Case study 2 that the semantic pre-processing stage, which was not a major concern in Case study 1, was an essential first step. The data from the bibliographic sources normally contains errors, such as misspelling the author's name, the journal title, or in the references list. Occasionally, additional information has to be added to the original data, for example, if the author's address is incomplete or wrong. For this reason, the analysis cannot be applied directly to the data retrieved from the bibliographic sources – a pre-processing stage over the retrieved data is necessary to overcome these issues.

In Case study 2, the clustering approach was used to cluster the data into groups sharing common characteristics, graph based visualisation and mutual information were used to validate the approach. Case study 2 focused on developing and extending the approach used in Case study 1, allowing the approach to expand into important new directions to make use of both the node attributes and links, in a way that will produce better results when working with anomalies.

Clusters are designed to classify observations, as anomalies should fall in regions of the data space where there is a small density of normal observations. The anomalies occur in 2 as a

cluster among the data, such observations are called *collective anomalies*, defined by Chandola et al. (2009) as follows: “The individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together, as a collection is anomalous.” Existing work on collective anomaly detection requires supporting relationships to connect the observations, such as sequential data, spatial data and graph data. Mutual information can be used to interpret collective anomalies. Mutual information can contribute to our understanding of anomalous features and help to identify links with anomalous behaviour. In Case study 2, mutual information was applied to interpret the semantics of the clusters. In cluster 5, for example, mutual information found no links amongst this group of nodes. This indicates *collective anomalies*, as zero mutual information between two random variables means that the variables are independent. Link mining considers data sets as a linked collection of interrelated objects and therefore it focuses on discovering explicit links between objects. Using mutual information allows us to work with objects without these explicit links. Cluster 5 contained documents, which had been selected as part of the co-citation data, but these documents were not themselves cited. Mutual information allowed us to examine the relationships between documents and to determine that some objects made use of self-citation meaning that they were regarded co-cited but did not connect to other objects. We also identified a community anomaly, where the edge is considered a relationship anomaly, because it connects two communities, which are usually not connected to one another. Mutual information provided information about the relationships between objects, which could not be inferred from a clustering approach alone. This additional information supports a semantic explanation of anomalies.

5.3 Summary

Case study 2 was developed to address the issues identified in Case study 1 and also allowed us to use mutual information to validate the visualisation graph. We used a real world data set where the anomalies were not known in advance and the data required pre-processing. We were able to show that the approach developed in Case Study 1 scaled to large data volumes and combined with semantic pre-processing, allowed us to work with noisy and inconsistent data. Mutual information supported a semantic interpretation of the clusters, as shown by the discussion of cluster 5.

Case study 2 involved data pre-processing which demonstrated the adapted CRISP-DM method to link mining. A number of transformations were carried out before the modelling stage it consisted of a real world data set where the anomalies were not known in advance. This is to establish whether the proposed approach would be valid if used with a data set where the anomalies and relationships were unknown to investigate how mutual information can be applied to interpret the semantics of the anomalies.

In Case study 2 there are more complex relationships among authors, which needed to be validated; this was not of major importance in Case study 1. The data for Case study 2 consisted of co-citations extracted from the Web Of Science, (WOS) which is a real word data. The size of data was limited by the download restriction from the Web Of Science. The actual data used include 1001 records consisting of the following fields: authors, year, volume, page, and type of journal. The mutual information has applied on the co-cited authors who appeared first in the record. This data was a richer data then Case study 1 because of the potential number of relationships between co-citation documents within a cluster and with respect to other clusters. Having clustered and then visualised the data and examined the resulting visualised graph and the underlying cluster through mutual information, we were able to determine that the results produced were valid, demonstrating that the approach can be used with complex real world data set.

6 Conclusion and future work

6.1 Introduction

Many real-world applications produce data which links to other data, such as the World Wide Web (hypertext documents connected through hyperlinks), social networks (such as people connected by friendship links) and bibliographic networks (nodes corresponding to authors, papers and the edges corresponding to cited-by). Link mining refers to data mining techniques that explicitly consider these links when building predictive or descriptive models of the linked data. Getoor and Diehl (2005), identify a set of commonly addressed link mining tasks which are: *Object-related tasks*, *Link-related tasks* (which has been used in Case study 1) and *Graph-related tasks* (which has been used in Case study 2). Recently there has been an exchange of ideas among these different approaches to link mining.

Link mining is an exciting and rapidly expanding area. The goal of this thesis was to show mutual information can help in providing a semantic interpretation of anomalies in the data, to characterise the anomalies, and how mutual information can help measuring the information that object item X shares with another object item Y. Whilst most link mining approaches focus on predicting link type, link based object classification or object identification, this research focused on using link mining to detect anomalies and discovering links/objects among anomalies. This thesis attempted to demonstrate the contribution of mutual information to interpret anomalies using two different case studies. The first Case study was used to test the approach and the second Case study was used to show its applicability to real data.

6.2 Evaluations of the main approach

The aim of this approach is to check data quality and any associated problems in order to discover first insights into the case studies, and detect interesting subsets to form hypotheses regarding hidden information. This approach can help to identify any anomalies in the data, to characterise them and to understand their properties. Mutual information is a quantitative measurement of how much one random variable (B) tells about another random variable (A). In this case, information is thought of as a reduction in the uncertainty of a variable; high mutual information indicates a large reduction in uncertainty whereas low mutual information

indicates a small reduction and zero mutual information between two random variables means that the variables are independent. The relationship between variables is essential to correctly determine whether the working values for the system. If A and B were identical, then all the information derived from obtaining variable A would supply the knowledge needed to get variable B. If two or more variables provide the same information or have similar effects on one outcome, this can be taken into consideration while constructing a model. Mutual information has been successful in detecting network intrusion (Gu *et al.*, 2006), self-propagating malicious codes detection (Khayam *et al.*, 2011) and mimicry attacks on host-based intrusion detection (Wagner & Soto, 2002).

In this thesis we considered the problem of detecting anomalies in two different types of datasets. The first Case study detected *point anomalies* and the second Case study identified *collective anomalies*. The method proposed in this thesis was evaluated first on a tightly constrained test data set and then on a real world data set. Evaluation of both data sets revealed that the proposed method tends to optimise the selection of candidates as anomalies. In chapter 4 we focused on a test data set, the sales datasets in Case study1. We started with the investigation of detecting individual record anomalies. In Case study 1 the aim was to determine mutual information could be used to explain the anomalies and the shared link. This method is especially useful when some of the attributes have a very high rarity, and when many of the attribute values are rare. We then considered the problem of detecting anomalous groups in data, which has been clustered using a hierarchical clustering approach. Chapter 5 described *collective anomalies*, which assumes that there is some self-similarity among the anomalous records, and that they are sufficiently anomalous to stand out by themselves. The experimental work—confirmed the effectiveness and efficiency of the proposed methods in practice. In particular, this revealed that our method is able to deal with data sets with a large number of objects and attributes. In Case study 2 having clustered and then visualised the data and examined the resulting visualisation graph and the underlying cluster through mutual information, we were able to determine that the results produced were valid, demonstrating that the approach can be used with the real world data set.

Anomalies detection finds applications in many domains, where it is desirable to determine interesting and unusual events in the activity, which generates such data. The core of all anomalies detection methods is the creation of a probabilistic, statistical or algorithmic

model, which characterises the normal behaviour of the data. The deviations from this model are used to determine the anomalies. A good domain-specific knowledge of the underlying data is often crucial in order to design simple and accurate models, which do not over fit the underlying data. Using mutual information contributes to our understanding of the anomalous features and helps with semantic interpretation and to identify links with anomalous behaviour. The problem of anomalies detection becomes especially challenging, when significant relationships exist among the different data points. This is the case for bibliographic data in which the patterns in the relationships among the data points play a key role in defining the anomalies.

In the data used in Case study 2, there is significantly more complexity in terms of how anomalies may be defined or modelled which can be used to interpret semantic meaning. In general, *the more complex the data, the more the analyst has to make prior inferences of what is considered normal for modelling purposes* (Aggarwal *et al.*, 2011). Therefore, anomalies may be defined in terms of significant changes in the underlying network community or distance structure. Such models combine network analysis and change detection in order to detect structural and temporal anomalies from the underlying data.

6.2.1 Finding of case study1

Case study1 used a two-step clustering setting. The measure of mutual information between two variables takes all association patterns into account when estimating the extent to which the two variables co-vary with each other. Therefore, this mutual information-based measure is a way of inferring links in data.

In Case study1, the transaction value of the dataset is the obvious and straightforward interpretation, as it contains very high and very low transaction values compared to the normal range of expenditure for that individual. Identifying groups of individuals or objects that have similar transaction values to each other, however yet they are different from individuals in other groups that can be distinctive, sufficient and have semantic features.

Our approach was applied to a small dataset where transactional data is structured data and patterns are stable. Context plays an important role in anomalies detection, because patterns used to detect anomalies cannot take into account all environmental factors, it is necessary to put each anomaly, once detected, in context. This information can be used to justify the behaviour of an object. This is another reason why good situational awareness is needed to

describe an event. Relevant contextual data qualify the anomaly detections (Seibert, 2009). The results from Case study1 have provided evidence of additional context anomalies such as point anomalies; this strongly suggests understanding the domain of information source it has an important role to play in anomaly detection.

The results using the algorithm are satisfactory with regard to proof of concept data, synthetic anomalies values and the small size of anomalies. This presents more assurance to the approach.

6.2.2 Finding of Case study 2

The co-citation data applied hierarchical clustering and visualised the data as a graph where nodes represented authors and edges represented cited-by. The aim was to cluster the nodes into groups sharing common characteristics; mutual information was applied to all clusters and demonstrated strong links among the element of each cluster, except in cluster 5. Mutual information conforms that cluster 5 elements share no links with the clusters and among themselves no link was found between authors. Zero mutual information between two random variables means that the variables are independent. As the discussion in chapter 5 shows mutual information can provide a semantic interpretation of anomalous features.

6.3 Research contributions

6.3.1 Major contributions

1. The study of anomalies in link mining

Link mining considers datasets as a linked collection of interrelated objects and therefore it focuses on discovering explicit links between objects. The proposed novel approach advocates the use of mutual information to identify vital hidden information in link mining applications. The proposed method is novel because it supports the semantic interpretation of anomalies in the context of research citation data. The research focuses on detecting anomalies in two different case studies, proof of concept data and co-citation data, using mutual information to the semantic interpretation. This research has adapted the method used in the emerging field of link mining. The challenge in forming a universal structure for anomaly detection is that the definitions of anomalies and normality are typically domain-specific. This has led to domain-specific efforts in this area based on factors such as the type

of anomaly, the nature of the data, the availability of data labels and other constraints. The approach developed in this Case study is illustrated with reference to bibliographic data but is not domain specific and can be applied in any context where the interpretation of anomalies is important.

2. Applying MI to provide semantic interpretation of anomalies

Applying mutual information contributes to our understanding of the anomalous features and helps identify links with anomalous behaviour. Data attributes (features) in anomalies detection are divided into two distinct groups: context (or condition) attributes B, and target attributes A. Anomalies detection attempts to interpret anomalies in target attributes A with respect to context attributes B. The contextual feature allows identification of patterns that are typically in one context but anomalous in the other. This has led to domain-specific efforts in this area based on factors such as the type of anomaly, the nature of the data, the availability of data labels, and other constraints. Mutual Information has been used in the context of text mining and data mining and also in link prediction. The novel contribution of this thesis is the development of an approach, which allows mutual information to be applied to provide a semantic interpretation of links in the context of link mining.

To apply mutual information to envisage new trends, of new emerging research area, or new community formation and many other application domains to improve the efficiency. The applicability of the approach to data sets of greater scale and diversity is a matter of future research such as security and health domains.

3. MI applied to validate clustering and visualisation.

In Case study 2, hierarchical clustering is applied to identify clusters and the data is visualised using graph representation. Anomalies occur as a cluster among the data, such observations are *collective anomalies*. Cluster validity with respect to anomalies can be difficult to evaluate because of data volumes. This research has demonstrated that mutual information can be applied to evaluate cluster content and the validity of the clustering approach. This also supports validation of the visualisation element.

6.3.2 Minor contributions

1. Modified CRISP to support link mining

There is not yet a comprehensive methodology that can support link mining tasks. The CRISP-DM process, which is well established methodology used by data mining researches, can provide a solid basis to support link mining tasks. This thesis has adapted CRISP-DM to support link-mining studies.

2. Applied CRISP to support link mining

The adapted CRISP-DM methodology, which consists of six stages, has been applied in this thesis these are: problem definition, data description, data pre-processing, data transformation, data exploration, data modelling and data evaluation. In the modelling stage in Case study 1 we used a two-step cluster and in Case study 2 we used graph mapping, hierarchical cluster and visualisation.

6.4 Limitations of the study

The thesis concludes by recognising certain limitations:

1. Time period: The co-citation data used was limited to an arbitrarily chosen period of time up to 2011.
2. Co-citation data was extracted from three databases SCI-EXPANDED, SSCI, A&HCI.
3. Uses a restricted subset of co-citation data and limited feature construction and analysis to first authors.

6.5 Challenges

A number of Challenges were faced in this thesis; these are outlined below:

1. Difficulty to identify suitable software to support the visualisation in Case study 2. The software employed was suitable for the data set used in Case study 2 but an alternative would be required for work with a larger or more complex data set to visualise data more clearly.
2. Data volumes and data quality presented a challenge in Case study 2 as the bibliographic data was noisy and needed cleaning, in terms of detecting misspelled and duplicate items;

there was a large number of items in the data that represented the same object or concept but with different spelling. In other cases, a concept was represented using different words (lexical forms) or acronyms, and yet referred to the same concept.

3. Feature construction is a great challenge. The study focused on basic object feature, such as first authors and cited-by. The link based approach would benefit from using attributes of these objects.

4. This is a fast evolving field; techniques and approaches have evolved during the course of the research. It has become a challenging task to keep up to date with the ever growing literature.

6.6 Future work

The current study can be extended in a variety of ways.

1. To extend the approach by working with a dynamic set of data, for example data related to dynamic social networks, scientific communities structures, detection of criminal communities.

2. To apply mutual information to support the prediction of anomalous links; mutual information can be combined with link prediction models in order to identify potential links to help develop strategies and policies. Prediction is an important part of decision-making in business, medicine and many other application domains to improve the efficiency of predictions.

3. To apply mutual information to predict trends, of new emerging research area, or new community formation.

4. Bibliometric graphs can be used to:

- Identify the main research areas in a scientific field, and gain insight on the size of the different domains.
- View how the areas link to each other.

- Can be used in a number of different contexts. Researchers can use bibliometric graph to get an overview of the field in which they are active or to perform a high-level
- Bibliometric graph can also be of value to scientific publishers, journal editors and librarians that may for example use these maps to explore how a journal is positioned relative to other related journals.
- Other possible applications of bibliometric a graph are in science teaching (e.g., Börner et al., 2009; Klavans & Boyack, 2009) and in the history, philosophy, and sociology of science (e.g., Small, 2003).

5. Mutual information has been used to provide a interpreted semantic between objects and the strength of the links, which can support the analysis and exploration of the data. For example in this study, we utilise feature selection for link mining, which are considered to influence citing behaviours. The idea is that link provides the tool to discover ‘anomalous links’, that are unexpected and therefore interesting. An unexpected citation in a paper citation network may be a sign of interdisciplinary working because the number of papers has increased and research areas have been segmented, it has become more difficult for both researchers and reviewers to decide which papers should be cited. Links that affect the existence and the class of links helps us make decisions, which will support citation even with a huge amount of data.

7 References

- Abe S., Kawano H., Goldstein J., Ohtani S., Solovyev S.I., Baishev D.G. and Yumoto K. (2006) Simultaneous identification of a plasmaspheric plume by a ground magnetometer pair and IMAGE Extreme Ultraviolet Imager. *Journal of Geophysical Research* 111(A11).
- Adami C.(2004) Information theory in molecular biology. *Physics of Life Reviews*.1.p.3–22.
- Aggarwal C., and Yu P.(2002). Redefining Clustering for High-Dimensional Applications. *Proceedings of the IEEE International Conference on Transaction of Knowledge and Data Engineering*. 14 (2). P.210 – 225.
- Aggarwal C., and Yu P.(2001) Outlier Detection for High Dimensional Data. *International Conference on Management of Data*. 30(2). P.37 – 46.
- Aggarwal R, Isil E, Miguel A. Ferreira, and Matos P.(2011) Does Governance Travel Around the World? Evidence from Institutional Investors, *Journal of Financial Economics* 100. P.154-181.
- Aggarwal R., Gehrke J., Gunopulos D.,and Raghavan P.(1998) Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 27(2). p.94 – 105.
- Aggarwal Y. Zhao, and Yu P.S.(2011) Outlier Detection in Graph Streams, *ICDE Conference*.
- Aghagolzadeh H. Soltanian-Zadeh B. Araabi, and Aghagolzadeh A.(2007) A hierarchical clustering based on mutual information maximization. *In IEEE ICIP*.p.277–280.
- Akoglu L., McGlohon M., Faloutsos C.,(2010) OddBall: Spotting Anomalies in Weighted Graphs. *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*.
- Alfuraih S., Sui N., and McLeod D .(2002) Using Trusted Email to Prevent Credit Card Frauds in Multimedia Products. *World Wide Web: Internet and Web Information Systems*, 5 (3). P.244 – 256.
- Allan J., Carbonell J., Doddington G., Yamron J., and Yang Y. (1998) Topic detection and tracking pilot study. *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*.
- Almansoori W., Gao S., Jarada T., Elsheikh A., M, Murshed A., Jida J., Alhaji R., and Rokne J.(2011) Link prediction and classification in social networks and its application in healthcare and systems biology. *Netw Model Anal Health Info Bioinformation* 1(1–2).p.27–36.
- Badia A., Kantardzic M.(2005) Link Analysis Tools for Intelligence and Counterterrorism. *ISI*. P.49-59.

Barbara D., Li Y., Couto J., Lin J. L., and Jajodia S.(2003) Bootstrapping a data mining intrusion detection system. *Proceedings of the 2003 ACM symposium on Applied computing*. ACM Press.

Barlow D. H.(1988) Anxiety and its disorders: The nature and treatment of anxiety and panic. New york, Guilford.

Bhattacharyya D.K., and Borah B.(2004) An Improved Sampling-based DBSCAN for Large Spatial Databases. *Proceedings of the International Conference on Intelligent Sensing and Information*. P.92.

Bindewald and Shapiro.(2006) indewald E, Shapiro BA: RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA*.12.p.342-352.

Bolton R. J. and Hand D. J.(1999) Unsupervised profiling methods for fraud detection. *In Conference on Credit Scoring and Credit Control 7, Edinburgh*.

Börner J., Mendoza A., & Vosti S. A.(2009) Ecosystem services, agriculture, and rural poverty in the Eastern Brazilian. *Amazon: Interrelationships and policy prescriptions. Ecological Economics*.64.p.356–373.

Brachman, R. J. & Anand, T., “The process of knowledge discovery in databases.”, AAAI Press / The MIT Press. 1996.

Brockett P. L., Xia, X., and Derrig R. A.(1998) Using kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. *Journal of Risk and Insurance*. 65(2) P.245-274.

Burns N., Grove SK.(2005) The Practice of Nursing Research: Conduct, Critique, and Utilization (5th Ed.). St. Louis, Elsevier Saunders.

Buslje C.M., et al.(2010) Networks of High Mutual Information Define the Structural.

Butte A, Kohane I.(2000) Unsupervised Knowledge Discovery in Medical Databases Using Relevance Networks. *Proc Amia Symp* .In Press.

Chai K.H., Ding Y. and Xing Y.(2009) Quality and Customer Satisfaction Spillovers in the Mobile Phone Industry Service Science.1(2).p.93-106.

Chakrabarti and Avik.(2001) The Determinants of Foreign Direct Investment: Sensitivity Analyses of Cross-Country Regressions. *Kyklos, Wiley Blackwell*.54(1).p.89-113.

Chandola V., Banerjee A., and Kumar V.(2009) Anomaly Detection. *A Survey, ACM Computing Survey*. 41(3). p.15.

Chandola V., Eilertson E., Ertoz L., Simon, G., and Kumar V. (2006) Data mining for cyber security. *Data Warehousing and Data Mining Techniques for Computer Security*, A. Singhal, Ed. Springer.

Chapman, P. et al, "CRISP-DM 1.0 - Step-by-step data mining guide." SPSS, 2000

Chau D. H., Pandit S., Faloutsos C. (2006) Detecting fraudulent 1032 personalities in networks of online auctioneers. In: *Knowledge Discovery in Databases: PKDD*. p.103–114.

Chellappa., Rama J., and Anil. (1993) *Boston: Academic Press*.

Chen Z., Hendrix W., Samatova N. F. (2012) Community-based anomaly detection in evolutionary networks. *Journal of Intelligent Information Systems*. 39(1). p.59–85.

Christopher D., Manning, Prabhakar R., and Hinrich S., (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Church K. and Hanks P. (1990) Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*. 16(1). p.22-29.

Cover T. M., and Thomas JA. (2006) *Elements of Information Theory* (2nd ed).

Cox C., Enno A., Deveridge S., Seldon M., Richards R., Martens V., and Woodford P. (1997) Remote electronic blood release system. *Transfusion*. 37. p.960-974.

Creamer, G., and Stolfo, S. (2009) A link mining algorithm for earnings forecast and trading Data. *Min Knowl Disc*. 18. P.419–445.

Creswell J. (2003) Research design: Qualitative, quantitative and mixed methods approaches (2nd ed.). *Thousand Oaks, CA: SAGE Publications*.

Creswell J. W. (1994) Research design: Qualitative and quantitative approaches. *Thousand Oaks, CA: SAGE Publications*.

Dash M., and LIU H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(1–4). P.131–156.

Date SV., Marcotte. (2003) EM: Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol*. 21. p.1055-1062.

Dawy Z., Goebel B. (2006) Gene mapping and marker clustering using Shannon's mutual information, *IEEE Trans. Oncomputational biology and bioinformatics*. 3(1).

Day W., & Edelsbrunner H., (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*. 1: 7–24.

- Dejordy R., Borgatti S.P., Roussin C & Halgin D.S.(2007) Visualizing proximity data. *Field Methods*.19.p.239–63.
- Desforges D. M., Lord C. G., Ramsey S. L.(1998) Effects of structured cooperative contact on changing negative attitudes toward stigmatized social groups. *Journal of Personality and Social Psychology*.60.p.531 -544.
- DesJardins M., and Matthew E.(2006) Gaston, Speaking of relations: Connecting statistical relational learning and multi-agent systems. *ICML Workshop on Open Problems in Statistical Relational Learning*, Pittsburgh, PA.
- Doan AH., Madhavan J., Domingos P., and Halevy A.(2004) Ontology matching: A machine learning approach. *Handbook on ontologies*.
- Duda R. O., Hart P. E., and. Stork D. G.(2000) Pattern Classification and Scene Analysis, *John Wiley & sons*.
- Eagle N. and Pentland A.(2006) Reality Mining: Sensing Complex Social Systems. *Personal and Ubiquitous Computing*. 10(4).p.255-268.
- Eagle N., Pentland A., and Lazer D.(2009) Inferring friendship network structure by using mobile phone data. *PNAS*.
- Eisen M. B. , Spellman P. T. , Brown P. O. & Botstein D.(1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA***95**.p.14863– 14868.
- Emamian V., Kaveh M., and Tewfik A.(2000) Robust clustering of acoustic emission signals using the kohonen network. *Proceedings of the IEEE International Conference of Acoustics,Speech and Signal Processing. IEEE Computer Society*.
- ErtÄoz A., Arnold M., Prerau L., Portnoy., and Stolfo S.(2003) A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. *In Proceedings of the Data Mining for Security Applications Workshop*.
- Ertoz L.; Steinbach, M.; Kumar V.(2004). Finding Topics in collections of documents: A shared nearest neighbour approach. *Clustering and Information Retrieval*. P.83-104.
- Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S.(2002) A geometric framework for unsupervised anomaly detection. *Proceedings of Applications of Data Mining in Computer Security. Kluwer Academics*.P.78-100.
- Ester M., Kriegel H-P., Sander J., and Xu X.(1996) A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. P.226 – 231.
- Farhoomand A.F.(1987) Scientific progress of management information systems. *The data base for Advances in Information Systems* 18(4).p.48–56.

Fawcett T., Provost F.(1999) Activity monitoring: noticing interesting changes in behavior. *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining (KDD-99)*.p.53–62.

Fayyad, U., Piatetsky-Shapiro G., Smyth P. (1996). *Advances in Knowledge Discovery and Data Mining*. MIT Press.

Feilong C., Scripps J., and Tan P.(2008) Link Mining for a Social Bookmarking Web Site.*In Proc of IEEE/WIC/ACM International Conference on Web Intelligence (WI-2008), Sydney, Australia.*

Fern X. Z., & Brodley C. E.(2003) Random projection for high dimensional data clustering: A cluster ensemble approach. *ICML*.

Fränti P., and Kivijärvi J.,(2000) Randomised Local Search Algorithm for the Clustering Problem. *Pattern Analysis and Applications*.3(4). p.358 – 369.

Fujimaki R., Yairi T., and Machida K.(2005) An approach to spacecraft anomaly detection problem using kernel feature space. *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM Press, New York.p.401–410.

Gao H.(2010). Aerobic denitrification in permeable Wadden Sea sediments. *ISME J*(4).p. 417–426.

Garfield E.(1980) Is Information Retrieval in the Arts and Humanities Inherently Different from that in Science. *Library Quarterly*. 50(1).p.40-57.

Gero J., and Saunders R.(2000) Constructed representations and their functions in computational models of designing, *Proceedings of the Fifth Conference on Computer Aided Architectural Design Research in Asia (CAADRIA 2000)*, CASA, Singapore. p.215–224.

Getoor L.(2003). Link mining. A new data mining challenge, *SIGKDD Explorations*, 5(1). p.84-89.

Getoor L.(2005) .Tutorial on Statistical Relational Learning. *ILP*: 415.

Getoor L., and Diehl C.(2005). Link mining: A survey *SIGKDD Explorations*, December. Vol.7 (2).

Ghosh S., and Reilly D. L.(1994). Credit card fraud detection with a neural-network. *Proceeding of the 27th Annual Hawaii International Conference on System Science*.3.

Goldenberg, A., Shmueli, G., Caruana, R. & Fienberg, S. (2002). Early Statistical Detection of Anthrax Outbreaks by Tracking Over-the-Counter Medication Sales. *Proc. of the National Academy of Sciences*, 5237-5249.

Grubbs Frank E.(1969) Procedures the data to assure that the results for detecting outlying observations in are representative of the thing samples. *Technometrics* 11.p.1-2.

Gu G.,Steeg. Ver ., and Galstyan A.(2013). Statistical Tests for Contagion in Observational Social Network Studies. *AISTATS'13*.

Gu G., Fogla P., Dagon D., Lee W., Skoric B.(2006). Measuring intrusion detection capability: An information-theoretic approach. *Proc of ACM Symposium on InformAction, Computer and Communications Security (ASIACCS)*.

Guha S., Rastogi R., and Shim K.(2001). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems* 25(5). p. 345-366.

Halkidi M., Batistakis Y.,and Vazirgiannis M.(2002) Cluster Validity Methods: part I. *Proceedings of the ACM SIGMOD International Conference on Management of Data.* (31) 2, p.40 – 45.

Han J., and Kamber M.(2001) Data Mining: Concepts and Techniques. *The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers.* P.550.

Harris T.(1993). Neural network in machine health monitoring. *Professional Engineering*.

Hastie T., Tibshirani R., Friedman J., (2009). Hierarchical clustering. *The Elements of Statistical Learning.*New York: Springer. 520–528.

Hawkins S., He H., Williams G. J., and Baxter R. A.(2002) Outlier detection using replicator neural networks. *In Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery.*p.170–180.

He Y., Vogelstein B., Velculescu VE., Papadopoulos N., Kinzler KW.(2008). *The antisense transcriptomes of human cells Science.* 322.p.1855–1857.

He Z., Deng S., and Xu X.(2002). Outlier detection integrating semantic knowledge. *Proceedings of the Third International Conference on Advances in Web-Age Information Management.*

He Z., Xu X., and Deng S.(2003). Discovering Cluster-based Local Outliers. *Pattern Recognition Letters.* 24(9-10). P.1641 – 1650.

He Z., Xu X., and Deng S.(2005) An optimization model for outlier detection in categorical data. *Proceedings of International Conference on Intelligent Computing.* 3644.

Hero A., Ma B., Michel O., and Gorman J.(2002a) Alpha-divergence for classification, indexing and retrieval. *Communications and Signal Processing Laboratory Technical Report CSPL-328.*

- Hlavackova-Schindler K., Paluš M., Vejmelka M., and Bhattacharya J.(2007) Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441.p.1–46.
- Hu T., and Sung S.Y.(2003) Detecting pattern-based outliers. *Pattern Recognition Letters*.24 (16). P.3059 – 3068.
- Hubert B., Milan S., Grocott A., Blockx C., Cowley S W A., and Gérard G.C.(2006). Dayside and nightside reconnection rates inferred from IMAGE FUV and Super Dual Auroral Radar Network data. *J. Geophys. Res.*, 111, A03217, doi:10.1029/2005JA011140.
- Hughes T.R., Marton M.J., Jones A.R., Roberts C.J., Stoughton R., Armour C.D., Bennett H.A., and Friend S.H.(2000) Functional discovery via a compendium of expression profiles. *Cell*.102.p.109–126.
- Hulle M., and Van M.(2008) Constrained subspace ICA based on mutual information optimization directly. *Neural Computation*.20.p.964–973.
- Jaing M., Tseng S., and Su C.(2001) Two-phase Clustering Process for Outlier Detection. *Pattern Recognition Letters*. 22(6–7). p.691–700.
- Kybic., J. (2006) Consistent and elastic registration of histological sections using vector-spline regularization *Arg approaches to medical image*.
- Jiang H., Huang Y., Zhuang Z. and Hwang K.C.(2001) Fracture in mechanism-based strain gradient plasticity. *Journal of Mechanics and Physics of Solids*.49.p.979–993.
- Jin W., Tung A., and Han J.(2001). Mining Top-n Local Outliers in Large Databases. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.p.293 – 298.
- Johnson R. B., and Onwuegbuzie A. J.(2004) Mixed methods research: A research paradigm whose time has come. *Educational Researcher*. 33(7).p.14-26.
- Johnson S. C., (1967): "Hierarchical Clustering Schemes" *Psychometrika*, 2:241-254.
- Kaufman L., and Rousseeuw P.J.(1990) Finding Groups in Data. *John Wiley & Sons, New York*.
- Keogh E., Lonardi S., and Ratanamahatana C. A.(2004). Towards parameter-free data mining. *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*. *ACM Press*.p. 206-215.
- Khayam S. A., Ashfaq A. B and Radha H.(2011) Joint network-host based malware detection using information-theoretic tools. *Journal in Computer Virology*. 7(2).p.159-172.
- Kirkland, D., Senator, T., Hayden, J., Dybala, T., Goldberg, H. & Shyr, P. (1999). The NASD Regulation Advanced Detection System. *AAAI* 20(1): Spring, 55-67.

Klavans R., and Boyack K. W.(2009) Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*.

Kohonen T.(1997) Self-organizing maps. *Springer-Verlag New York, Inc*.

Kraskov A., and Grassberger P.(2009) MIC: Mutual information based hierarchical clustering. In F. Emmert-Streib & M. Dehmer (Eds.). *Information theory and statistical learning* (p. 101-123). *Springer US*.

Kraskov A., Stogbauer H., Andrzejak R. G., and Grassberger P.(2005) Hierarchical clustering using mutual information. *Europhysics Letters*.70(2).p.278.

Kuang L., and Zulkernine M.(2008) An anomaly intrusion detection method using the CSI-KNN algorithm. *SAC*.P. 921-926.

Labib K., and RaoVemuri V.(2002) “NSOM: A Real-time Network-Based Intrusion detection System Using Self-Organizing Maps, *Networks and Security*.

Lakhina A., Crovella M., and Diot C.(2005) Mining Anomalies Using Traffic Feature Distributions. *Proceedings of ACM SIGCOM*.p. 217-228.

Lee W., and Stolfo, S.(2000) A framework for constructing features and 638 models for intrusion detection systems. *ACM Transactions on Information and System Security*. 3(4).

Leonenko N., Pronzato L and Savani V.(2008a) A class of Renyi information estimators ´ for multidimensional densities. *Annals of Statistics*.36(5).p.2153–2182.

Lewi J., Butera R., and Paninski L.(2007) Real-time adaptive information-theoretic optimization of neurophysiology experiments. *Advances in Neural Information Processing Systems*.19.

Li P., Kenneth W., and Church.(2007) A sketch algorithm for estimating two-way and multi-way associations. *Computational Linguistics*.33(3).p.305–354.

Li W.(1990) Mutual information functions versus correlation functions. *Journal of statistical physics*.

Li., Ping., Church and Kenneth W.(2007) A sketch algorithm for estimating two-way and multi-way associations. *Computational Linguistics (Preliminary results appeared in HLT/EMNLP 2005)*. 33(3).p.305–354.

Liben-Nowell D., and Kleinberg J.(2003) The link prediction problem for social networks. In CIKM ’03. *Proceedings of the twelfth international conference on Information and knowledge management*.p.556–559.

Lin H., Fan W., and Wallace L.(2007) An empirical study of web-based knowledge community success. *Proceedings of the 40th Hawaii International Conference on System Sciences*. P.1530-160.

Lin S., and Brown D.(2004) An Outlier-based Data Association Method for Linking Criminal Incidents. *Proceedings of the SIAM International Conference on Data Mining*.

Lin S., and Brown D.(2003) An Outlier-based Data Association Method. *Proceedings of the SIAM International Conference on Data Mining*.

Mahoney M. V., Chan P. K., and Arshad M. H.(2003) A machine learning approach to anomaly detection. Tech. *Department of Computer Science, Florida Institute of Technology Melbourne FL 32901*.

Manning C., Prabhakar R., and Schütze H.(2008) Introduction to Information Retrieval. *Cambridge University Press*.

Marchette D.(1999) A statistical method for profiling network traffic. *Proceedings of 1st USENIX Workshop on Intrusion Detection and Network Monitoring*. Santa Clara, CA, .p.119-128.

Marcus G., Fernandes K., and Johnson S.(2007) Infant rule-learning facilitated by speech. *Psychol. Sci.*18.p.387–391.

Markou M., and Sing S.(2003) Novelty Detection: a review- parts 1 and 2. *Signal Processing*, 83(12).P.2481-2521.

Meila M.(2005). Comparing clusterings: an ax-iomatic view. *Proceedings of the 22nd international conference on Machine learning*.p. 577-584.

Miller E., and Fisher III J.(2003) ICA using spacings estimates of entropy. *JMLR*.4.p.1271–1295.

Miller E., Narayana., and Hanson A.(2013) Coherent motion segmentation in moving camera videos using optical flow orientations. *ICCV*.

Motulsky H.(1995), Intuitive Biostatistics .*Oxford University Press*, New York, 386.0-1950-8607-4.

Mustafa Y. T., Tolpekin V., and Stein A., and Sub M.(2007) The application of Expectation Maximization algorithm to estimate missing values in Gaussian Bayesian network modeling for forest growth. *IEEE Transactions on Geoscience and Remote Sensing*.

Newman M. E. J., and Girvan M.(2003) Finding and evaluating community structure in network. *Phys. Rev. E* 69.

Noyons E.(2001) Bibliometric mapping of science in a science policy context. *Scientometrics*, 50(1).p.83-98.

O'Madadhain J., Smyth P., and Adamic L.(2005) Learning Predictive Models for Link Formation. *To be presented at the International Sunbelt Social Network Conference*.

- Olle P.(2010). Are highly cited papers more international?. *Scientometrics*.83(2).p.397-401.
- Otey M. E., Ghoting A., and Parthasarathy S.(2003) Fast distributed outlier detection in mixed-attribute data sets. *Data Mining and Knowledge Discovery*. 12(2-3) p.203-228.
- Otey M., Parthasarathy S., Ghoting A., Li G., Narravula S., and Panda D.(2003).
- P'oczozs Z., and L'orincz A.(2009) Complex independent process analysis. *PLA University of Science & Technology, Nanjing 210007, China*.
- Brunel N., Panzeri S., Logothetis NK., and Kayser C.(2010) Sensory neural codes using multiplexed temporal scales. *Trends Neurosci*.33.p.111–120.
- Patcha A., and Park JM.(2007) An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*. 51(12).p.3448-3470.
- Pedreschi, D.(2008) Mining on Complex (Social) Network. *Pisa: at: didawiki.cli.di.unipi.it/lib/exe/fetch.php/.../wma.sna.pedreschi.3.pdf*.
- Peng H., Long F., and Ding C.(2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*.27(8).p.1226-38.
- Petrovskiy M.(2003) Outlier Detection Algorithms in Data Mining Systems. *Programming and Computing Software*. 29(4).p.228 – 237.
- Phua C., Alahakoon D., and Lee V.(2004) Minority Report in Fraud Detection Classification of Skewed Data. *Special Issue on Learning from Imbalanced Datasets*. 6(1).p.50 – 59.
- Pires A., and Santos-Pereira C.(2005) Using clustering and robust estimators to detect outliers in multivariate data. *Proceedings of International Conference on Robust Statistics. Finland*.
- Piatetsky-Shapiro, G., and Matheus, C. 1994. The Interestingness of Deviations. In *Proceedings of KDD-94*, eds. U. M. Fayyad and R. Uthurusamy. Technical Report WS-03. Menlo Park, Calif.: AAAI Press.
- Platt J.(2000) Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. A. *Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, Eds*.p.61–74.
- Poczozs B., and L'orincz A.(2005) Independent subspace analysis using geodesic spanning trees. *ICML*. p.673–680.
- Poczozs B., and L'orincz A.(2009) Identification of recurrent neural networks by Bayesian interrogation techniques. *Journal of Machine Learning Research*.10.p.515–554.

- Poczos B., and Lorincz A.(2005) Cross-entropy optimization for independent process analysis. *In Proceedings of Independent Component Analysis and Blind Signal Separation (ICA 2006)*.3889.p.909–916.
- Popescul A., Ungar L., Lawrence S., and Pennock D.(2003) Statistical re-lational learning for document mining. *Computer and Information Sciences, University of Pennsylvania*.
- Price D., and de Solla J.(1965) Networks of scientific papers. *Science*.149. p.510- 515.
- Provana K.G., Leischowc S. J., Keagyb J., and Nodorac J.(2010) Research collaboration in the discovery, development, and delivery networks. *of a statewide cancer coalition*.33(4).p. 349-355.
- Ramadas M., Ostermann S., and Tjaden B. C.(2003) Detecting anomalous network traffic with self-organizing maps. *Proceedings of Recent Advances in Intrusion Detection*.P.36-54.
- Ramaswamy S., Rastogi R., and Shim K.(2000). Efficient Algorithms for Mining Outliers from Large Data Sets. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. 29(2).p.427 – 438.
- Rattigan M. J., and Jensen D.(2005) The case for anomalous link discovery. *SIGKDD Explorations*, 7(2).
- Santos, M & Azevedo, C (2005). Data Mining – Descoberta de Conhecimento em Bases de Dados. FCA Publisher.
- SAS Enterprise Miner – SEMMA. SAS Institute, 2014 [online] available: <http://www.sas.com/technologies/analytics/datamining/miner/semma.html> (September 2014.)
- Savage D., Zhang X., Yu X., Chou P., and Wang Q.(2014) Anomaly Detection in Online Social Networks. *Social Networks*.39.p.62–70.
- Scarth G., McIntyre M., Wowk B., and Somorjai R.(1995) Detection of novelty in functional images using fuzzy clustering. *Proceedings of the 3rd Meeting of International Society for Magnetic Resonance in Medicine. Nice, France*.p.238.
- Seibert J.(2009) Land-cover impacts on streamflow: A change-detection modeling approach that incorporates parameter uncertainty. *Hydrological Sciences Journal, in press*.
- Senator T. E. (2005) Link mining applications: progress and challenges, *ACM SIGKDD Explorations Newsletter*.7(2).p.76-83.
- Sequeira K. and Zaki M.(2002) Admit: anomaly-based data mining for intrusions. *In Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press*.p.386–395.

Shafique, U.Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*. 12 (1), 217-222.

Shan Y., Sawhney H., and Kumar R.(2005) Vehicle Identification between Non-Overlapping Cameras without Direct Feature Matching. Computer Vision. *IEEE International Conference on I*.

Shannon C.E.(1948). A mathematical theory of communication. *Bell System Technical Journal*.27.p. 379–423 and 623–656.

Shearer C., The CRISP-DM model: the new blueprint for data mining, *J Data Warehousing* (2000); 5:13—22.

Seretan V., Wehrli E. (2006). Accurate collocation extraction using a multilingual parser. In Proceedings of *COLING/ACL*.

Sheikholeslami G., Chatterjee S., and Zhang A.(1998) Wavecluster: A multi-resolution clustering approach for very large spatial databases. *Proceedings of the 24rd International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc.p.428-439.

Shyu M.L., Chen S.C., Sarinnapakorn K., and Chang L.(2003) A novel anomaly detection scheme based on principal component classifier. *Proceedings of 3rd IEEE International Conference on Data Mining*.p.353–365.

Skillicorn D. B.(2004) Detecting Related Message Traffic, Workshop on Link Analysis, Count ErtÄoz errorism, and Privacy. *SIAM International Conference on Data Mining*, Seattle, USA.

Small H.(1973) Co-citation in the scientific literature: A new measurement of the relationship between two documents. *Journal of the American Society of information science and technology*.24(4).p.265-269.

Small H.(2003) Paradigms, citations, and maps of science: A personal history. *Journal of the American Society for Information Science and Technology*.54(5).p.394-399.

Smith R., Bivens A., Embrechts M., Palagiri C., and Szymanski B.(2002) Clustering approaches for anomaly based intrusion detection. *Proceedings of Intelligent Engineering Systems through Artificial Neural Networks*. ASME Press.P.579-584.

Song X., Wu M., Jermaine C., and Ranka S.(2007) Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*.19(5).p.631-645.

Sparrow M.(1991) The application of network analysis to criminal intelligence: an assessment of the prospects. *Soc Netw* 13.p.251–274.

Steuer R., Kurths J., Daub C.O., Weise J., and Selbig J.(2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*. 18(2).p.S231–S240.

Strehl A., and Ghosh J.(2002) Relationship-based clustering and visualization for high-dimensional data mining. *INFORMS Journal on Computing*. *in press*.

Su, P., Shang C., and Shen A.(2013) Soft Computing - A Fusion of Foundations, Methodologies and Applications archive. 17(12).p.2399-2410.

Tan L., Tanir D., and Smith K.(2005) Introduction to Data Mining. *Addison-Wesley*.J(2).p.229-245.

Taskar B., Abbeel P., and Koller D.(2003) Discriminative probabilistic models for relational data. *Proc. UAI02, Edmonton, Canada*.

Theodoridis S., and Koutroubas K.(1999) Pattern Recognition. *Academic Press*.

Thottan., and Ji.(2003) Anomaly detection in IP networks. *Signal Processing, IEEE Transactions* .51(8).p.2191-2204.

Tomovic A., and Oakeley EJ.(2007) Position dependencies in transcription factor binding sites.*Bioinformatics*.23.p.933–941.

Tumer K., and Agogino A.K.(2008) Adaptive Management of Air Traffic Flow: A Multiagent Coordination Approach. *AAAI*.p.1581-1584.

Umair S and Haseeb Q. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*. 12 (1), 217-222.

Van Eck N.J and Waltman L.(2009a) VOSviewer: A computer program for bibliometric mapping. Technical Report ERS-2009-005-LIS, Erasmus University Rotterdam. Erasmus research institute of mangment. Available at: <http://hdl.handle.net/1765/14841>.

Van Eck N.J., and Waltman L.(2009b) VOSviewer: A computer program for bibliometric mapping. In B. Larsen and J. Leta, editors, *Proceedings of the 12th International Conference on Scientometrics and Informetrics*.p.886–897.

Vinueza A., and Grudic G.(2004) Unsupervised outlier detection and semi-supervised learning.Tech. Rep. CU-CS-976-04, Univ. of Colorado at Boulder.

Wadhah,A.,Gao,S., Jarada,T., Elsheikh,A.,Murshed,A.,Jida,J.,Alhajj,R.,on Rokne. (2011). Link prediction and classification in social networksand its application in healthcare and systems biology. *Netw Model Anal Health Inform Bioinforma*. 1 (1), 27-36

Wagner D., and Soto P.(2002). Mimicry Attacks on Host-Based Intrusion Detection Systems. *ACM CCS*.

Wanjantuk P., and Keane J.A.(2004) Finding related documents via communities in the citation graph. *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on*.1.p.445-450.

Watts D. J., and Strogatz S. H.(1998) Collective dynamics of 'small-world' networks". *Nature* **393** (6684).p.440–442.

Waugh C. K., & Ruppel M.(2004) Citation analysis of dissertations, thesis and research paper references in workforce education and development. *The Journal of Academic Librarianship*.30(4).p.276-284.

Wei L., Qian W., Zhou A., and Jin W.(2003). Hot: Hypergraph-based outlier test for categorical data. *Proceedings of the 7th Pacic-Asia Conference on Knowledge and Data Discovery*. p.399-410.

Williams C. B.(2005) The lived experiences of women in executive positions of the United States federal civil service. D.M. *dissertation, University of Phoenix, United States* . Publication No. AAT 3202470).

Williams G., Baxter R., He H., Hawkins S., and Gu L.(2002) A Comparative Study for RNN for Outlier Detection in Data Mining. *Proceedings of the 2nd IEEE International Conference on Data Mining*.p.709.

Wu N. , and Zhang J.(2003) Factor analysis based anomaly detection. *Proceedings of IEEE Workshop on Information Assurance. United States Military Academy, West Point, NY*.

Wu J., Xiong H., and Chen J.(2009) .Adapting the right measures for k-means clustering, in *KDD*.p.877–886.

Xu, K.M., Zhang M, Eitzen Z.A., Ghan S.J., Klein S.A., and Zhang J.(2005) Modeling springtime shallow frontal clouds with cloud-resolving and single-column models. *J. Geophys. Res.*, 110, D15S04, doi:10.1029/2004JD005153.

Yankov D., Keogh E. J., and Rebbapragada U.(2007). Disk aware discord discovery: Finding unusual time series in terabyte sized datasets. *Proceedings of International Conference on Data Mining*.p.381-390.

Yang Y., Zhiguo G., and Leong H.U.(2011). Identifying points of interest by self-tuning clustering. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11)*. ACM, New York.

Ypma A., and Duin R.(1998). Novelty detection using self-organizing maps. *Progress in Connectionist Based Information Systems*.2.p.1322-1325.

Yu D., Sheikholeslami G., and Zhang A.(2002) Findout: finding outliers in very large datasets. *Knowledge And Information Systems*.4(4).p. 387.

Zilin Z, Rui Z & Youliang Z. (2014). A Hybrid Feature Selection Method Based on Rough Conditional Mutual Information and Naive Bayesian Classifier. *ISRN Applied Mathematics*. 2014 (1), 1-11

Glossary

Actor: Actor refers to a person, organisation, or nation that is involved in a social relation. Hence, an actor is a vertex in a social network.

Adjacent: Two vertices are adjacent if they are connected by a line.

Arc: An arc is a directed line. Formally, an arc is an ordered pair of vertices.

Anomalies: Something that deviates from what is standard, normal, or expected (Chandola *et al*, 2009).

Anomalies detection: to detecting patterns in a given data set those do not conform to an established normal behaviour (Chandola *et al*, 2009).

Clique: A clique is a maximal complete subnetwork containing: three vertices or more.

Cluster-Based Local Outlier Factor (CBLOF): A measure for identifying the physical of an outlier is designed.

Degree: The degree of a vertex is the number of lines incident with it.

Density: Density is the number of lines in a simple network, expressed as a proportion of the maximum possible number of lines.

Dyad: A dyad is a pair of vertices and the lines between them.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN): is a data clustering algorithm.

Edge: An edge is an undirected line. Formally, an edge is an unordered pair of vertices.

Indegree: The indegree of a vertex is the number of arcs it receives.

Expectation–maximization algorithm (EM): is an iterative method for finding maximum likelihood or maximum a posterior (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables.

FindCBLOF: cluster based local discovering outliers.

Link: in this research a link refers to some real world connection between two entities (Senator 2005).

Link mining (LM): technique that explicitly considers links when building predictive or descriptive models of the linked data (Getoor & Diehl 2005).

Mutual information (MI): is one of many quantities that measure the reduction in uncertainty about one random variable given knowledge of another by the application of data (Gray, 1990).

Neighbour: A vertex that is adjacent to another vertex is its neighbour.

Node: In a network, a node is a connection point, either a redistribution point or an end point for data transmissions.

Noises: are random errors or variance in measured variables, and should be removed before outlier's detection (Chandola *et al*, 2009).

Outliers: are observations that are numerically distant from the rest of the data (Chandola *et al*, 2009).

Relation: A relation is collection of specific ties among members of a group.

RObust Clustering using linKs (ROCK): clustering algorithm for categorical and Boolean attributes.

Shared Near Neighbour graph (SNN): clustering algorithm for shared near neighbour in graph.

Self-organizing map (SOM): is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional, discredited representation of the input space of the training samples, called a map.

Partition: A partition of a network is a classification or clustering of the vertices in the network such that each vertex is assigned to exactly one class or cluster.

Two-mode network: In a two-mode network, vertices are divided into two sets and vertices can be related only to vertices in the other set.

Undirected graph: An undirected graph contains no arcs: all of its lines are edges.

Vertex (vertices): A vertex (singular of vertices) is the smallest unit in a network.

Weakly connected: A network is weakly connected if each pair of vertices is connected by a semipath.

Appendix A

Using MATLAB to calculate the mutual information between three attributes.

```
%=====

echo on;

a = [1 2 2 3 5 1 2 1 3 3 1 1 3 3 3 4 4 4 4 3 4]';
b = [1 1 1 1 1 1 1 2 2 3 5 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000]';
c = [3 3 3 3 3 5 5 1 2 2 2 1 3 3 3 3 3 3 3 4 4]';

mutualinfo(a,b)
mutualinfo(a,c)
mutualinfo(b,c)

entropy(a)
entropy(b)
entropy(c)

jointentropy(a,b)
jointentropy(a,c)
jointentropy(b,c)

condmutualinfo(a,b,c)
condmutualinfo(a,c,b)
condmutualinfo(b,c,a)

mutualinfo(a,b,c)

entropy(a)+condentropy(b)-jointentropy(a,b)
condentropy(a,b)
condentropy(a,c)
jointentropy(a,c)
mutualinfo(a,c)
condmutualinfo(a,c)
condmutualinfo(a,c,b)
condmutualinfo(a,c,[b c])

echo off;

function h = mutualinfo(vec1,vec2)
%

[p12, p1, p2] = estpab(vec1,vec2);
h = estmutualinfo(p12,p1,p2);

function h = entropy(vec1)
if nargin<1,

    disp('Usage: h = entropy(vec1).');
```

```

    h = -1;

else,

    [p1] = estpa(vec1);
    h = estentropy(p1);

end;

function h = condentropy(vec1,vec2)

if nargin<1,

    disp('Usage: h = condentropy(vec1,<vec2>).');
    h = -1;

elseif nargin<2,

    [p1] = estpa(vec1);
    h = estentropy(p1);

else

    [p12, p1, p2] = estpab(vec1,vec2);
    h = estcondentropy(p12,p2);

end;

function h = jointentropy(vec1,vec2)
%=====

if nargin<1,

    disp('Usage: h = condentropy(vec1,<vec2>).');
    h = -1;

elseif nargin<2,

    [p1] = estpa(vec1);
    h = estentropy(p1);

else,

    [p12] = estpab(vec1,vec2);
    h = estjointentropy(p12);

end;

function h = condmutualinfo(vec1,vec2,condvec)

if nargin<3,
    condvec = [];
end;

```

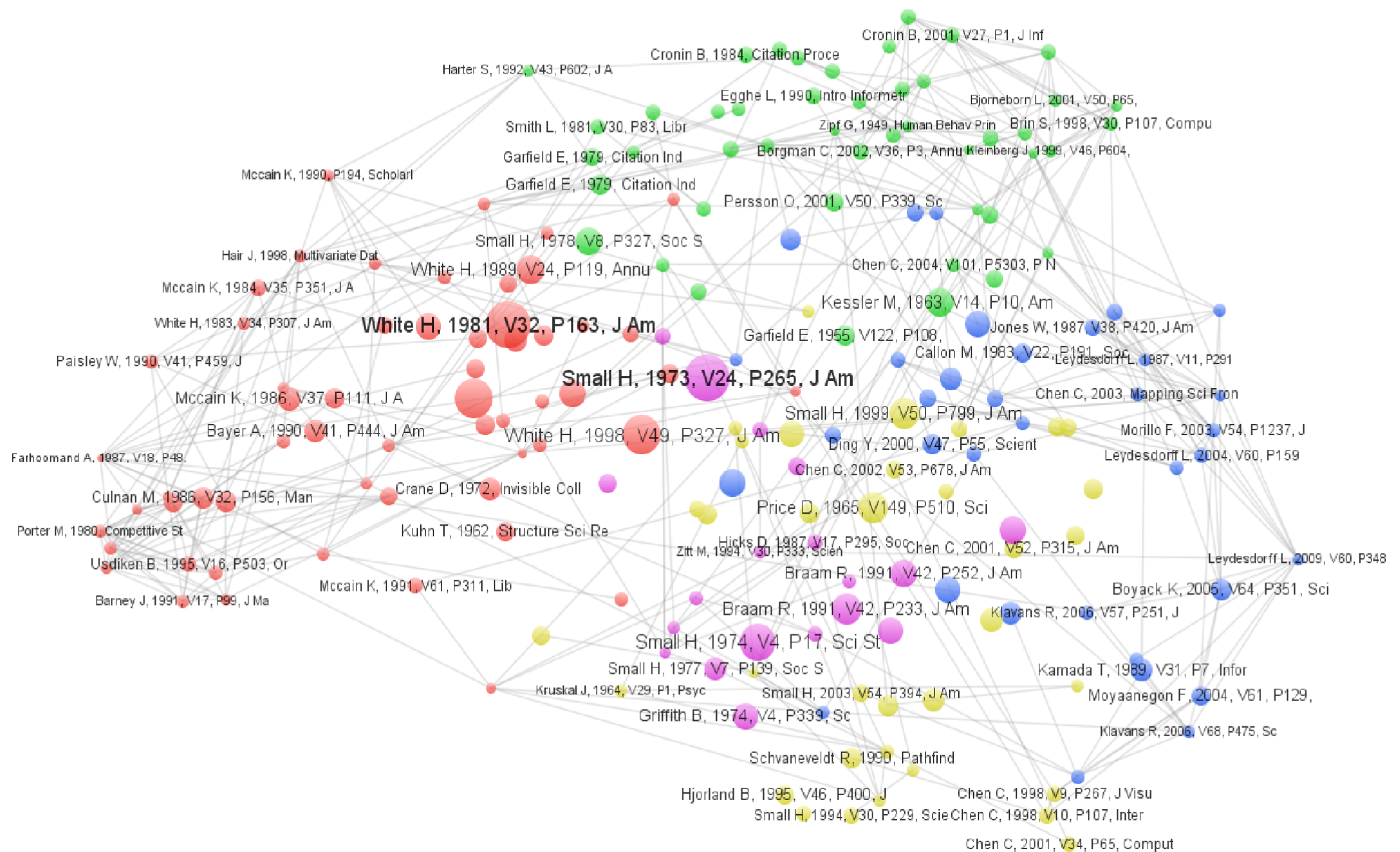
```

if size(condvec,2)>1,
    newcondvec_z = mergemultivariables(condvec);
else %including the case of condvec=[]
    newcondvec_z = condvec;
end;

if isempty(newcondvec_z),
    h = condentropy(vec2) - condentropy(vec2,vec1);
else
    newcondvec_xz = mergemultivariables(newcondvec_z,vec1);
    h = condentropy(vec2,newcondvec_z) - condentropy(vec2,newcondvec_xz);
end;

```


Appendix B: Case study 2 mapping of nodes (visualisation of nodes)



Appendix B

Case study 2: list of vertices (nodes)

*Vertices 193

- 1 "McCain K, 1990, V41, P433, J Am Soc Inform Sci"
- 2 "White H, 1981, V32, P163, J Am Soc Inform Sci"
- 3 "Small H, 1973, V24, P265, J Am Soc Inform Sci" 4
- "Small H, 1974, V4, P17, Sci Stud"
- 5 "White H, 1998, V49, P327, J Am Soc Inform Sci"
- 6 "Kessler M, 1963, V14, P10, Am Doc"
- 7 "Braam R, 1991, V42, P233, J Am Soc Inform Sci"
- 8 "Griffith B, 1974, V4, P339, Sci Stud"
- 9 "Braam R, 1991, V42, P252, J Am Soc Inform Sci"
- 10 "Culnan M, 1986, V32, P156, Manage Sci"
- 11 "Culnan M, 1987, V11, P341, Mis Quart"
- 12 "Price D, 1965, V149, P510, Science"
- 13 "White H, 1989, V24, P119, Annu Rev Inform Sci"
- 14 "Small H, 1978, V8, P327, Soc Stud Sci"
- 15 "Small H, 1985, V7, P391, Scientometrics"
- 16 "Small H, 1985, V8, P321, Scientometrics"
- 17 "Bayer A, 1990, V41, P444, J Am Soc Inform Sci"
- 18 "Persson O, 1994, V45, P31, J Am Soc Inform Sci"
- 19 "White H, 1990, P84, Scholarly Communicat"
- 20 "Ahlgren P, 2003, V54, P550, J Am Soc Inf Sci Tec, Doi 10.1002/Asi.10242"
- 21 "Small H, 1999, V50, P799, J Am Soc Inform Sci"
- 22 "McCain K, 1991, V42, P290, J Am Soc Inform Sci"
- 23 "McCain K, 1986, V37, P111, J Am Soc Inform Sci"
- 24 "Small H, 1977, V7, P139, Soc Stud Sci"
- 25 "White H, 2003, V54, P1250, J Am Soc Inf Sci Tec, Doi 10.1002/Asi.10325"
- 26 "White H, 1982, V38, P255, J Doc"
- 27 "Marshakova I, 1973, V2, P3, Nauchno Tekhnicheskaya"
- 28 "Culnan M, 1990, V41, P453, J Am Soc Inform Sci"
- 29 "White H, 1997, V32, P99, Annu Rev Inform Sci"
- 30 "Small H, 1980, V2, P277, Scientometrics"
- 31 "Leydesdorff L, 2006, V57, P1616, J Am Soc Inf Sci Tec"
- 32 "Chen C, 1999, V35, P401, Inform Process Manag"
- 33 "White H, 2003, V54, P423, J Am Soc Inf Sci Tec, Doi 10.1002/Asi.10228"
- 34 "Ding Y, 1999, V25, P67, J Inform Sci"
- 35 "Garfield E, 1979, Citation Indexing"
- 36 "Small H, 1986, V37, P97, J Am Soc Inform Sci" 37
- "McCain K, 1984, V35, P351, J Am Soc Inform Sci" 38
- "Garfield E, 1964, Use Citation Data Wr"
- 39 "White H, 1981, V32, P16, J Am Soc Inform Sci"
- 40 "Karki R, 1996, V22, P323, J Inform Sci"
- 41 "Garfield E, 1955, V122, P108, Science"
- 42 "Small H, 1980, V36, P183, J Doc"

43 "Price D, 1963, Little Sci Big Sci"

44 "Callon M, 1983, V22, P191, Soc Sci Inform"

45 "Persson O, 2001, V50, P339, Scientometrics"

46 "Crane D, 1972, Invisible Coll Diffu"

47 "Kamada T, 1989, V31, P7, Inform Process Lett"

48 "Boyack K, 2005, V64, P351, Scientometrics"

49 "Moya-anegon F, 2004, V61, P129, Scientometrics"

50 "Culnan M, 1986, V10, P289, Mis Quart"

51 "Callon M, 1986, Mapping DynamicsSci"

52 "Bensman S, 2004, V55, P935, J Am Soc Inf Sci Tec, Doi 10.1002/Asi.20028"

53 "Small H, 1993, V26, P5, Scientometrics"

54 "Garfield E, 1972, V178, P471, Science"

55 "Borner K, 2003, V37, P179, Annu Rev Inform Sci"

56 "Salton G, 1983, Intro Modern Informa"

57 "Small H, 1997, V38, P275, Scientometrics"

58 "Leydesdorff L, 1987, V11, P295, Scientometrics"

59 "Kuhn T, 1962, Structure Sci Revolu"

60 "Cottrill C, 1989, V11, P181, Knowledge"

61 "Cronin B, 2001, V27, P1, J Inform Sci"

62 "Small H, 1985, V11, P147, J Inform Sci"

63 "Chen C, 2001, V52, P315, J Am Soc Inf Sci Tec"

64 "Paisley W, 1990, V41, P459, J Am Soc Inform Sci"

65 "Ramosrodriguez A, 2004, V25, P981, Strategic Manage J"

66 "Wasserman S, 1994, Social Network Anal"

67 "Noyons E, 1999, V50, P115, J Am Soc Inform Sci"

68 "Schvaneveldt R, 1990, Pathfinder Ass Netwo"

69 "Almind T, 1997, V53, P404, J Doc"

70 "White H, 1986, V5, P93, Inform Technol Libr"

71 "Brin S, 1998, V30, P107, Comput Networks Isdn"

72 "Astrom F, 2007, V58, P947, J Am Soc Inf Sci Tec, Doi 10.1002/Asi.20567"

73 "Gmur M, 2003, V57, P27, Scientometrics"

74 "McCain K, 1990, P194, Scholarly Communicat"

75 "Eom S, 1996, V47, P941, J Am Soc Inform Sci"

76 "McCain K, 1998, V41, P389, Scientometrics"

77 "Leydesdorff L, 2004, V60, P371, J Doc, Doi 10.1108/00220410410548144"

78 "Leydesdorff L, 2004, V60, P159, Scientometrics"

79 "Kuhn T, 1970, Structure Sci Revolu"

80 "Chen C, 2006, V57, P359, J Am Soc Inf Sci Tec, Doi 10.1002/Asi.20317"

81 "Lin X, 1997, V48, P40, J Am Soc Inform Sci"

82 "Sullivan D, 1977, V7, P223, Soc Stud Sci"

83 "Peters H, 1995, V46, P9, J Am Soc Inform Sci"

84 "Macroberts M, 1989, V40, P342, J Am Soc Inform Sci"

85 "Noyons E, 2001, V50, P83, Scientometrics"

86 "Small H, 2003, V54, P394, J Am Soc Inf Sci Tec, Doi 10.1002/Asi.10225"

87 "Ding Y, 2000, V47, P55, Scientometrics"

88 "Leydesdorff L, 2005, V56, P769, J Am Soc Inf Sci Tec"

89 "White H, 1983, V34, P307, J Am Soc Inform Sci"

90 "Kruskal J, 1978, Multidimensional Sca"

91 "Borgman C, 2002, V36, P3, Annu Rev Inform Sci"

92 "Hjorland B, 1995, V46, P400, J Am Soc Inform Sci"

93 "Moravcsik M, 1975, V5, P86, Soc Stud Sci"

94 "Chen C, 1998, V10, P107, Interact Comput"

95 "Chen C, 1998, V9, P267, J Visual Lang Comput"

96 "Pritchard A, 1969, V25, P348, J Doc"

97 "Salton G, 1983, Intro Modern Inform"

98 "McCain K, 1991, V61, P311, Libr Quart"

99 "Boyack K, 2002, V53, P764, J Am Soc Inf Sci Tec"

100 "Hicks D, 1987, V17, P295, Soc Stud Sci"

101 "Zhao D, 2006, V42, P1578, Inform Process Manag"

102 "Egghe L, 1990, Intro Informetrics Q"

103 "Borgman C, 1990, Scholarly Communicat"

104 "Pilkington A, 1999, V19, P7, Int J Oper Prod Man"

105 "Hoffman D, 1993, V19, P505, J Consum Res"

106 "Usdiken B, 1995, V16, P503, Organ Stud"

107 "Small H, 1979, V1, P445, Scientometrics"

108 "Kleinberg J, 1999, V46, P604, J Acm"

109 "Kruskal J, 1964, V29, P1, Psychometrika"

110 "Garfield E, 1979, Citation Indexing It"

111 "Hair J, 1998, Multivariate Data An"

112 "Fruchterman T, 1991, V21, P1129, Software Pract Exper"

113 "Mullins N, 1977, V42, P552, Am Sociol Rev"

114 "Perry C, 1998, V49, P151, J Am Soc Inform Sci"

115 "Vargasquesada B, 2007, Visualizing Structur"

116 "Melin G, 1996, V36, P363, Scientometrics"

117 "Klavans R, 2006, V68, P475, Scientometrics"

118 "Leydesdorff L, 2009, V60, P348, J Am Soc Inf Sci Tec"

119 "Klavans R, 2006, V57, P251, J Am Soc Inf Sci Tec, Doi 10.1002/Asi.20274"

120 "Noyons E, 1998, V49, P68, J Am Soc Inform Sci"

121 "Morillo F, 2003, V54, P1237, J Am Soc Inf Sci Tec, Doi 10.1002/Asi.10326"

122 "Bjorneborn L, 2001, V50, P65, Scientometrics"

123 "Eom S, 1996, V16, P315, Decis Support Syst"

124 "Porter M, 1980, Competitive Strategy"

125 "Porter M, 1985, Competitive Advantag"

126 "Osareh F, 1996, V46, P149, Libri"

127 "Osareh F, 1996, V46, P217, Libri"

128 "Small H, 1999, V48, P72, Libr Trends"

129 "Callon M, 1991, V22, P155, Scientometrics"

130 "Moyaanegon F, 1998, V42, P229, Scientometrics"

131 "Small H, 1994, V30, P229, Scientometrics"
 132 "Edge D, 1979, V17, P102, Hist Sci"
 133 "Zitt M, 1994, V30, P333, Scientometrics"
 134 "Jones W, 1987, V38, P420, J Am Soc Inform Sci"
 135 "Ingwersen P, 1998, V54, P236, J Doc"
 136 "Chubin D, 1975, V5, P423, Soc Stud Sci"
 137 "Chen C, 2001, V34, P65, Computer"
 138 "Healey P, 1986, V15, P233, Res Policy"
 139 "Farhoomand A, 1987, V18, P48, Data Base"
 140 "Deerwester S, 1990, V41, P391, J Am Soc Inform Sci"
 141 "Cronin B, 1984, Citation Process Rol"
 142 "Chen C, 2002, V53, P678, J Am Soc Inf Sci Tec"
 143 "Larson R, 1996, P71, P 59 Ann M Am Soc In"
 144 "Mccain K, 1983, V5, P277, Scientometrics"
 145 "Morris S, 2003, V54, P413, J Am Soc Inf Sci Tec, Doi
 10.1002/Asi.10227"
 146 "Merton R, 1973, Sociology Sci Theore"
 147 "Salton G, 1979, V22, P146, Ieee T Prof Commun"
 148 "Smith L, 1981, V30, P83, Libr Trends"
 149 "Morris T, 1998, V5, P448, J Am Med Inform Assn"
 150 "Reader D, 2006, V30, P417, Entrep Theory Pract"
 151 "Schildt H, 2006, V30, P399, Entrep Theory Pract"
 152 "Pinski G, 1976, V12, P297, Information Processi"
 153 "White H, 2001, V52, P87, J Am Soc Inf Sci Tec"
 154 "Borgatti S, 2002, Ucinet Windows Softw"
 155 "Leydesdorff L, 2006, V57, P601, J Am Soc Inf Sci Tec"
 156 "Price D, 1976, V27, P292, J Am Soc Inform Sci"
 157 "Leydesdorff L, 1987, V11, P291, Scientometrics"
 158 "Andrews J, 2003, V91, P47, J Med Libr Assoc"
 159 "Tsay M, 2003, V57, P7, Scientometrics"
 160 "Griffith B, 1972, V177, P959, Science"
 161 "Barney J, 1991, V17, P99, J Manage"
 162 "Lotka A, 1926, V16, P317, J Washington Academy"
 163 "Leydesdorff L, 1989, V18, P209, Res Policy"
 164 "Gilbert G, 1977, V7, P113, Soc Stud Sci"
 165 "Merton R, 1968, V159, P56, Science"
 166 "Leydesdorff L, 1998, V43, P5, Scientometrics"
 167 "Leydesdorff L, 1997, V48, P418, J Am Soc Inform Sci"
 168 "Harter S, 1992, V43, P602, J Am Soc Inform Sci"
 169 "Cohen W, 1990, V35, P128, Admin Sci Quart"
 170 "Bush V, 1945, V176, P101, Atlantic Monthly"
 171 "Chubin D, 1976, V17, P448, Sociological Q"
 172 "Moed H, 2005, Citation Anal Res Ev"
 173 "Peters H, 1993, V22, P23, Res Policy"
 174 "Lievrouw L, 1989, V16, P615, Commun Res"
 175 "Mccain K, 1995, V46, P306, J Am Soc Inform Sci"

176 "Chen C, 2003, Mapping Sci Frontier"
 177 "Garfield E, 1963, V14, P289, Am Doc"
 178 "Rip A, 1984, V6, P381, Scientometrics"
 179 "Rice R, 1988, V15, P256, Hum Commun Res"
 180 "Price D, 1970, P3, Communication Sci En"
 181 "Small H, 1981, V17, P39, Information Processi"
 182 "King J, 1987, V13, P261, J Inform Sci"
 183 "Zipf G, 1949, Human Behav Principl"
 184 "Swanson D, 1987, V38, P228, J Am Soc Inform Sci"
 185 "Vanraan A, 1990, V347, P626, Nature"
 186 "Narin F, 1976, Evaluative Bibliomet"
 187 "Freeman L, 1979, V1, P215, Soc Networks"
 188 "Chen C, 2004, V101, P5303, P Natl Acad Sci U S1"
 189 "Lawrence S, 1999, V32, P67, Ieee Comput"
 190 "Egghe L, 2002, V55, P349, Scientometrics"
 191 "Hirsch J, 2005, V102, P16569, P Natl Acad Sci Usa"
 192 "Hummon N, 1989, V11, P39, Soc Networks"
 193 "Gibbons M, 1994, New Production Knowl"
 *Edges
 1 2 77
 3 4 75
 3 2 73
 2 5 65
 1 5 55
 6 3 52
 7 3 43
 8 4 41
 3 5 39
 7 9 37
 1 3 34
 4 2 32
 10 11 32

Appendix C

The following provides more information about this algorithm using MATLAB:

```
*
* Implements Agglomerative Hierarchical Clustering algorithm.
*/
#include <float.h>
#include <math.h>
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#define NOT_USED 0 /* node is currently not used */
#define LEAF_NODE 1 /* node contains a leaf node */
#define A_MERGER 2 /* node contains a merged pair of root clusters */
#define MAX_LABEL_LEN 16
#define AVERAGE_LINKAGE 'a' /* choose average distance */
#define CENTROID_LINKAGE 't' /* choose distance between cluster centroids */
#define COMPLETE_LINKAGE 'c' /* choose maximum distance */
#define SINGLE_LINKAGE 's' /* choose minimum distance */
#define alloc_mem(N, T) (T *) calloc(N, sizeof(T))
#define alloc_fail(M) fprintf(stderr, \
                             "Failed to allocate memory for %s.\n", M)
#define read_fail(M) fprintf(stderr, "Failed to read %s from file.\n", M)
#define invalid_node(I) fprintf(stderr, \
                                "Invalid cluster node at index %d.\n", I)
typedef struct cluster_s cluster_t;
typedef struct cluster_node_s cluster_node_t;
typedef struct neighbour_s neighbour_t;
typedef struct item_s item_t;
float (*distance_fptr)(float **, const int *, const int *, int, int);
typedef struct coord_s {
    float x, y;
} coord_t;
struct cluster_s {
    int num_items; /* number of items that was clustered */
    int num_clusters; /* current number of root clusters */
    int num_nodes; /* number of leaf and merged clusters */
    cluster_node_t *nodes; /* leaf and merged clusters */
    float **distances; /* distance between leaves */
};
struct cluster_node_s {
    int type; /* type of the cluster node */
    int is_root; /* true if cluster hasn't merged with another */
    int height; /* height of node from the bottom */
    coord_t centroid; /* centroid of this cluster */
    char *label; /* label of a leaf node */
    int *merged; /* indexes of root clusters merged */
    int num_items; /* number of leaf nodes inside new cluster */
}
```

```

    int *items; /* array of leaf nodes indices inside merged clusters */
    neighbour_t *neighbours; /* sorted linked list of distances to roots */
};

struct neighbour_s {
    int target; /* the index of cluster node representing neighbour */
    float distance; /* distance between the nodes */
    neighbour_t *next, *prev; /* linked list entries */
};

struct item_s {
    coord_t coord; /* coordinate of the input data point */
    char label[MAX_LABEL_LEN]; /* label of the input data point */
};

float euclidean_distance(const coord_t *a, const coord_t *b)
{
    return sqrt(pow(a->x - b->x, 2) + pow(a->y - b->y, 2));
}

void fill_euclidean_distances(float **matrix, int num_items,
                             const item_t items[])
{
    for (int i = 0; i < num_items; ++i)
        for (int j = 0; j < num_items; ++j) {
            matrix[i][j] =
                euclidean_distance(&(items[i].coord),
                                  &(items[j].coord));
            matrix[j][i] = matrix[i][j];
        }
}

float **generate_distance_matrix(int num_items, const item_t items[])
{
    float **matrix = alloc_mem(num_items, float *);
    if (matrix) {
        for (int i = 0; i < num_items; ++i) {
            matrix[i] = alloc_mem(num_items, float);
            if (!matrix[i]) {
                alloc_fail("distance matrix row");
                num_items = i;
                for (i = 0; i < num_items; ++i)
                    free(matrix[i]);
                free(matrix);
                matrix = NULL;
                break;
            }
        }
        if (matrix)
            fill_euclidean_distances(matrix, num_items, items);
    } else
        alloc_fail("distance matrix");
    return matrix;
}

```

```

float single_linkage(float **distances, const int a[],
                    const int b[], int m, int n)
{
    float min = FLT_MAX, d;
    for (int i = 0; i < m; ++i)
        for (int j = 0; j < n; ++j) {
            d = distances[a[i]][b[j]];
            if (d < min)
                min = d;
        }
    return min;
}

float complete_linkage(float **distances, const int a[],
                      const int b[], int m, int n)
{
    float d, max = 0.0 /* assuming distances are positive */;
    for (int i = 0; i < m; ++i)
        for (int j = 0; j < n; ++j) {
            d = distances[a[i]][b[j]];
            if (d > max)
                max = d;
        }
    return max;
}

float average_linkage(float **distances, const int a[],
                     const int b[], int m, int n)
{
    float total = 0.0;
    for (int i = 0; i < m; ++i)
        for (int j = 0; j < n; ++j)
            total += distances[a[i]][b[j]];
    return total / (m * n);
}

float centroid_linkage(float **distances, const int a[],
                      const int b[], int m, int n)
{
    return 0; /* empty function */
}

float get_distance(cluster_t *cluster, int index, int target)
{
    /* if both are leaves, just use the distances matrix */
    if (index < cluster->num_items && target < cluster->num_items)
        return cluster->distances[index][target];
    else {
        cluster_node_t *a = &(cluster->nodes[index]);
        cluster_node_t *b = &(cluster->nodes[target]);
        if (distance_fptr == centroid_linkage)
            return euclidean_distance(&(a->centroid),
                                     &(b->centroid));
    }
}

```

```

        else return distance_fptr(cluster->distances,
                                   a->items, b->items,
                                   a->num_items, b->num_items);
    }
}

void free_neighbours(neighbour_t *node)
{
    neighbour_t *t;
    while (node) {
        t = node->next;
        free(node);
        node = t;
    }
}

void free_cluster_nodes(cluster_t *cluster)
{
    for (int i = 0; i < cluster->num_nodes; ++i) {
        cluster_node_t *node = &(cluster->nodes[i]);
        if (node->label)
            free(node->label);
        if (node->merged)
            free(node->merged);
        if (node->items)
            free(node->items);
        if (node->neighbours)
            free_neighbours(node->neighbours);
    }
    free(cluster->nodes);
}

void free_cluster(cluster_t * cluster)
{
    if (cluster) {
        if (cluster->nodes)
            free_cluster_nodes(cluster);
        if (cluster->distances) {
            for (int i = 0; i < cluster->num_items; ++i)
                free(cluster->distances[i]);
            free(cluster->distances);
        }
        free(cluster);
    }
}

void insert_before(neighbour_t *current, neighbour_t *neighbours,
                  cluster_node_t *node)
{
    neighbours->next = current;
    if (current->prev) {
        current->prev->next = neighbours;
        neighbours->prev = current->prev;
    }
}

```

```

        } else
            node->neighbours = neighbours;
        current->prev = neighbours;
    }
void insert_after(neighbour_t *current, neighbour_t *neighbours)
{
    neighbours->prev = current;
    current->next = neighbours;
}
void insert_sorted(cluster_node_t *node, neighbour_t *neighbours)
{
    neighbour_t *temp = node->neighbours;
    while (temp->next) {
        if (temp->distance >= neighbours->distance) {
            insert_before(temp, neighbours, node);
            return;
        }
        temp = temp->next;
    }
    if (neighbours->distance < temp->distance)
        insert_before(temp, neighbours, node);
    else
        insert_after(temp, neighbours);
}
neighbour_t *add_neighbour(cluster_t *cluster, int index, int target)
{
    neighbour_t *neighbour = alloc_mem(1, neighbour_t);
    if (neighbour) {
        neighbour->target = target;
        neighbour->distance = get_distance(cluster, index, target);
        cluster_node_t *node = &(cluster->nodes[index]);
        if (node->neighbours)
            insert_sorted(node, neighbour);
        else
            node->neighbours = neighbour;
    } else
        alloc_fail("neighbour node");
    return neighbour;
}
cluster_t *update_neighbours(cluster_t *cluster, int index)
{
    cluster_node_t *node = &(cluster->nodes[index]);
    if (node->type == NOT_USED) {
        invalid_node(index);
        cluster = NULL;
    } else {
        int root_clusters_seen = 1, target = index;
        while (root_clusters_seen < cluster->num_clusters) {
            cluster_node_t *temp = &(cluster->nodes[--target]);

```



```

        if (temp->type == NOT_USED) {
            invalid_node(index);
            cluster = NULL;
            break;
        }
        if (temp->is_root) {
            ++root_clusters_seen;
            add_neighbour(cluster, index, target);
        }
    }
}
return cluster;
}

#define init_leaf(cluster, node, item, len) \
do { \
    strncpy(node->label, item->label, len); \
    node->centroid = item->coord; \
    node->type = LEAF_NODE; \
    node->is_root = 1; \
    node->height = 0; \
    node->num_items = 1; \
    node->items[0] = cluster->num_nodes++; \
} while (0)

cluster_node_t *add_leaf(cluster_t *cluster, const item_t *item)
{
    cluster_node_t *leaf = &(amp;cluster->nodes[cluster->num_nodes]);
    int len = strlen(item->label) + 1;
    leaf->label = alloc_mem(len, char);
    if (leaf->label) {
        leaf->items = alloc_mem(1, int);
        if (leaf->items) {
            init_leaf(cluster, leaf, item, len);
            cluster->num_clusters++;
        } else {
            alloc_fail("node items");
            free(leaf->label);
            leaf = NULL;
        }
    } else {
        alloc_fail("node label");
        leaf = NULL;
    }
    return leaf;
}

#undef init_leaf
cluster_t *add_leaves(cluster_t *cluster, item_t *items)
{
    for (int i = 0; i < cluster->num_items; ++i) {
        if (add_leaf(cluster, &items[i]))

```

```

        update_neighbours(cluster, i);
    else {
        cluster = NULL;
        break;
    }
}
return cluster;
}

void print_cluster_items(cluster_t *cluster, int index)
{
    cluster_node_t *node = &(cluster->nodes[index]);
    fprintf(stdout, "Items: ");
    if (node->num_items > 0) {
        fprintf(stdout, "%s", cluster->nodes[node->items[0]].label);
        for (int i = 1; i < node->num_items; ++i)
            fprintf(stdout, ", %s",
                    cluster->nodes[node->items[i]].label);
    }
    fprintf(stdout, "\n");
}

void print_cluster_node(cluster_t *cluster, int index)
{
    cluster_node_t *node = &(cluster->nodes[index]);
    fprintf(stdout, "Node %d - height: %d, centroid: (%5.3f, %5.3f)\n",
            index, node->height, node->centroid.x, node->centroid.y);
    if (node->label)
        fprintf(stdout, "\tLeaf: %s\n\t", node->label);
    else
        fprintf(stdout, "\tMerged: %d, %d\n\t",
                node->merged[0], node->merged[1]);
    print_cluster_items(cluster, index);
    fprintf(stdout, "\tNeighbours: ");
    neighbour_t *t = node->neighbours;
    while (t) {
        fprintf(stdout, "\n\t\t%2d: %5.3f", t->target, t->distance);
        t = t->next;
    }
    fprintf(stdout, "\n");
}

void merge_items(cluster_t *cluster, cluster_node_t *node,
                cluster_node_t **to_merge)
{
    node->type = A_MERGER;
    node->is_root = 1;
    node->height = -1;
    /* copy leaf indexes from merged clusters */
    int k = 0, idx;
    coord_t centroid = { .x = 0.0, .y = 0.0 };
    for (int i = 0; i < 2; ++i) {

```

```

        cluster_node_t *t = to_merge[i];
        t->is_root = 0; /* no longer root: merged */
        if (node->height == -1 ||
            node->height < t->height)
            node->height = t->height;
        for (int j = 0; j < t->num_items; ++j) {
            idx = t->items[j];
            node->items[k++] = idx;
        }
        centroid.x += t->num_items * t->centroid.x;
        centroid.y += t->num_items * t->centroid.y;
    }
    /* calculate centroid */
    node->centroid.x = centroid.x / k;
    node->centroid.y = centroid.y / k;
    node->height++;
}

#define merge_to_one(cluster, to_merge, node, node_idx) \
do { \
    node->num_items = to_merge[0]->num_items + \
        to_merge[1]->num_items; \
    node->items = alloc_mem(node->num_items, int); \
    if (node->items) { \
        merge_items(cluster, node, to_merge); \
        cluster->num_nodes++; \
        cluster->num_clusters--; \
        update_neighbours(cluster, node_idx); \
    } else { \
        alloc_fail("array of merged items"); \
        free(node->merged); \
        node = NULL; \
    } \
} while(0)

cluster_node_t *merge(cluster_t *cluster, int first, int second)
{
    int new_idx = cluster->num_nodes;
    cluster_node_t *node = &(amp;cluster->nodes[new_idx]);
    node->merged = alloc_mem(2, int);
    if (node->merged) {
        cluster_node_t *to_merge[2] = {
            &(cluster->nodes[first]),
            &(cluster->nodes[second])
        };
        node->merged[0] = first;
        node->merged[1] = second;
        merge_to_one(cluster, to_merge, node, new_idx);
    } else {
        alloc_fail("array of merged nodes");
        node = NULL;
    }
}

```

```

    }
    return node;
}
#undef merge_to_one
void find_best_distance_neighbour(cluster_node_t *nodes,
                                int node_idx,
                                neighbour_t *neighbour,
                                float *best_distance,
                                int *first, int *second)
{
    while (neighbour) {
        if (nodes[neighbour->target].is_root) {
            if (*first == -1 ||
                neighbour->distance < *best_distance) {
                *first = node_idx;
                *second = neighbour->target;
                *best_distance = neighbour->distance;
            }
            break;
        }
        neighbour = neighbour->next;
    }
}

int find_clusters_to_merge(cluster_t *cluster, int *first, int *second)
{
    float best_distance = 0.0;
    int root_clusters_seen = 0;
    int j = cluster->num_nodes; /* traverse hierarchy top-down */
    *first = -1;
    while (root_clusters_seen < cluster->num_clusters) {
        cluster_node_t *node = &(cluster->nodes[--j]);
        if (node->type == NOT_USED || !node->is_root)
            continue;
        ++root_clusters_seen;
        find_best_distance_neighbour(cluster->nodes, j,
                                    node->neighbours,
                                    &best_distance,
                                    first, second);
    }
    return *first;
}

cluster_t *merge_clusters(cluster_t *cluster)
{
    int first, second;
    while (cluster->num_clusters > 1) {
        if (find_clusters_to_merge(cluster, &first, &second) != -1)
            merge(cluster, first, second);
    }
    return cluster;
}

```

```

}
#define init_cluster(cluster, num_items, items) \
do { \
    cluster->distances = \
        generate_distance_matrix(num_items, items); \
    if (!cluster->distances) \
        goto cleanup; \
    cluster->num_items = num_items; \
    cluster->num_nodes = 0; \
    cluster->num_clusters = 0; \
    if (add_leaves(cluster, items)) \
        merge_clusters(cluster); \
    else \
        goto cleanup; \
} while (0) \
cluster_t *agglomerate(int num_items, item_t *items)
{
    cluster_t *cluster = alloc_mem(1, cluster_t);
    if (cluster) {
        cluster->nodes = alloc_mem(2 * num_items - 1, cluster_node_t);
        if (cluster->nodes)
            init_cluster(cluster, num_items, items);
        else {
            alloc_fail("cluster nodes");
            goto cleanup;
        }
    } else
        alloc_fail("cluster");
    goto done;
cleanup:
    free_cluster(cluster);
    cluster = NULL;
done:
    return cluster;
}
#undef init_cluster
int print_root_children(cluster_t *cluster, int i, int nodes_to_discard)
{
    cluster_node_t *node = &(cluster->nodes[i]);
    int roots_found = 0;
    if (node->type == A_MERGER) {
        for (int j = 0; j < 2; ++j) {
            int t = node->merged[j];
            if (t < nodes_to_discard) {
                print_cluster_items(cluster, t);
                ++roots_found;
            }
        }
    }
}

```

```

        return roots_found;
    }
void get_k_clusters(cluster_t *cluster, int k)
{
    if (k < 1)
        return;
    if (k > cluster->num_items)
        k = cluster->num_items;
    int i = cluster->num_nodes - 1;
    int roots_found = 0;
    int nodes_to_discard = cluster->num_nodes - k + 1;
    while (k) {
        if (i < nodes_to_discard) {
            print_cluster_items(cluster, i);
            roots_found = 1;
        } else
            roots_found = print_root_children(cluster, i,
                                                nodes_to_discard);

        k -= roots_found;
        --i;
    }
}
void print_cluster(cluster_t *cluster)
{
    for (int i = 0; i < cluster->num_nodes; ++i)
        print_cluster_node(cluster, i);
}
int read_items(int count, item_t *items, FILE *f)
{
    for (int i = 0; i < count; ++i) {
        item_t *t = &(items[i]);
        if (fscanf(f, "%[^|]| %10f %10f\n",
                    t->label, &(t->coord.x),
                    &(t->coord.y)))
            continue;
        read_fail("item line");
        return i;
    }
    return count;
}
int read_items_from_file(item_t **items, FILE *f)
{
    int count, r;
    r = fscanf(f, "%5d\n", &count);
    if (r == 0) {
        read_fail("number of lines");
        return 0;
    }
    if (count) {

```

```

        *items = alloc_mem(count, item_t);
        if (*items) {
            if (read_items(count, *items, f) != count)
                free(items);
        } else
            alloc_fail("items array");
    }
    return count;
}

void set_linkage(char linkage_type)
{
    switch (linkage_type) {
        case AVERAGE_LINKAGE:
            distance_fptr = average_linkage;
            break;
        case COMPLETE_LINKAGE:
            distance_fptr = complete_linkage;
            break;
        case CENTROID_LINKAGE:
            distance_fptr = centroid_linkage;
            break;
        case SINGLE_LINKAGE:
        default: distance_fptr = single_linkage;
    }
}

int process_input(item_t **items, const char *fname)
{
    int count = 0;
    FILE *f = fopen(fname, "r");
    if (f) {
        count = read_items_from_file(items, f);
        fclose(f);
    } else
        fprintf(stderr, "Failed to open input file %s.\n", fname);
    return count;
}

int main(int argc, char **argv)
{
    if (argc != 4) {
        fprintf(stderr, "Usage: %s <input file> <num clusters> "
            "<linkage type>\n", argv[0]);
        exit(1);
    } else {
        item_t *items = NULL;
        int num_items = process_input(&items, argv[1]);
        set_linkage(argv[3][0]);
        if (num_items) {
            cluster_t *cluster = agglomerate(num_items, items);
            free(items);

```

```

        if (cluster) {
            fprintf(stdout, "CLUSTER HIERARCHY\n"
                "-----\n");
            print_cluster(cluster);
            int k = atoi(argv[2]);
            fprintf(stdout, "\n\n%d CLUSTERS\n"
                "-----\n", k);
            get_k_clusters(cluster, k);
            free_cluster(cluster);
        }
    }
    return 0;
}

```


Appendix D

Summary of literature on link mining and link mining techniques

No	Year	What	Who	Purpose	Technique/ methods	Tasks/challenge	PbS	Application area	Future work
1	2001	Frequent sub graph discovery	Michihiro & George	Finding frequent sub graph in large graph databases.	Association rules/Frequent Sub Graphs (FSG)	Sub graphs	Size of a transaction.	Graph isomorphism	Discover recurrent patterns in scientific, spatial, and relational datasets.
2	2003	Link prediction in relational data	Tasker et al.	Predicting the existence and the type of links between entities in domains. WebPages, a social network	Relational Markov Network (RMN)	The collective classification	cannot represent sub graph patterns	Universal Web pages & social works	Identify & predict objects interaction.
3	2003.	Link-based Classification using Labeled and Unlabeled Data	Lu & Getoor	look at some of the unique ways in which unlabeled data can improve performance when doing <i>link-based</i> classification,	Collective classification,	Link-based classification	-----	Citation	To use all of the information that unlabeled data provides.

Appendix D

Summary of literature on link mining and link mining techniques

4	2003	Statistical relational learning for link prediction	Popescul & Unger	Application for SRL to building link prediction regression models.	Statistical relation learning (SRL)	Link prediction	Standard statistical models, usually assume one table domain representation, which is inadequate for this task.	Scientific literature citations	Use intelligent search heuristics to speed up the discovery of subspaces with more useful features.
5	2003	Link-based Text Classification	Lu & Getoor	A statistical framework for modeling link distribution	logistic regression model	Link-based statistical models	Link statistic is not enough to capture the dependence.	Bibliographic dataset.	Using the link structure to help improve classification accuracy.
6	2003	Link mining :A new data mining challenge	Getoor	Summary of work & challenges in link mining and multi-relational data mining is coherently handling two different types of dependence.	-----	-----	A few Learning tasks range from predictive tasks, such as classification, to descriptive tasks, such as the discovery of frequently occurring sub-patterns.	Web, hypertext mining, mining social networks, security and law enforcement data, bibliographic citations and epidemiological.	Link mining is a promising new area where relational learning meets statistical modeling.

Appendix D

Summary of literature on link mining and link mining techniques

7	2004	Relational link based ranking	Geerts et al	Generalising link analysis methods for analyzing relational databases.	Random walk& The mutual reinforcement technique of HITS.	Link- based Rank	-----	Relational database and set of queries a unique weighted directed graph, which call the database graph.	How can the database graph be used to define measures of similarity between categorical data? Possible measures include the shortest path between tuples and the commute distance between nodes on the database graph.
8	2004	Deduplication and group detection using links.	Bhattacharya&getoor	how can be used to solve two entity deduplication and group discovery.	Clustering algorithms	Link-based clustering	An algorithm based just on entity attributes.	citation	How comparisons of the different distance measures for varying data characteristics that highlight the tradeoffs involved and results that show significant improvement over algorithms based just on entity attributes.
9	2005	Link mining for the semantic web position statement.	Getoor&.licamele	To develop ML algorithms.	Statistical machine learning for heterogeneous, linked data.	Link –based statistical models.	The meaning of a hyperlink between two resources on the internet cannot be understood by computers.	Semantic Web	use machine learning techniques which make use of ontological constraints together with the inferred semantic links&.learning the different kinds of links that exists.

Appendix D

Summary of literature on link mining and link mining techniques

10	2005	Prediction and ranking algorithms for event-based network data.	Joshua, et al	To study the problems of temporal link prediction and node ranking, and describe new methods that illustrate opportunities for data mining and machine learning techniques.	Markov random fields (MRFs)	Link prediction /link –based rank.	Time series	Social network analysis	new practical applications and for a better understanding of the dynamics of the underlying phenomena.
11	2005	Multi-relational data mining 2005 workshop report.	blocked& saso dzeroski	Finding patterns in expressive languages from multi-relational, complex and/or structured data.	ILP,KDD,ML	-----	Structured data	On multirelational and structural problems irrespective of origin and community.	-----
12	2005	Link Mining Applications Progress and Challenges	Senator	application and requirements in the area of complex event detection	-----	-----	There is not yet a comprehensive framework that can support a combination of link mining tasks as needed for many real applications	-----	-----

Appendix D

Summary of literature on link mining and link mining techniques

13	2005	An application of boosting to graph classification	Taku et al.	Application of boosting for classifying labeled graphs.	boosting /Kernel methods	graph classification	It is based on random walks in a graph.	Real world data such as chemical compounds, natural language texts, and bio sequences.	classification tasks involving discrete structural features
14	2005	privacy-enhanced linking	Sweeney	providing privacy protection within link analysis and introduces the notion of “privacy-enhanced linking”	link analysis/ privacy-enhanced linking(PEL)	collective consolidation	PEL privacy statement does not actually provide privacy but is consistent with minimizing same kinds of harms.	Guarantees and privacy protection	-----
15	2005	Discovery information connection sub graphs in multi-relational graphs.	Pamakrishn anet al	introduce heuristics that guide a subgraph discovery algorithms.	Display □-graph generation algorithm	A subgraph discovery	To develop the tools for finding correlation between patterns.	RDFGraphs	Algorithm development to support queries (RDF).
16	2005	Link mining: A survey	Getoor&, Christopher	Cover the core challenges addressed by a majority of ongoing research in the field.	-----	-----	Heterogeneous data sets	-----	-----

Appendix D

Summary of literature on link mining and link mining techniques

17	2005	relevance search anomaly detection in bipartite graph	Sun et al	Propose algorithms to compute the relevance score for each node.	Random walk	Subgraph discovery/anomaly link detection	Relevance search. Anomaly detection.	Collaborative Filtering	Predict users behavior and not anomalies
18	2006	An Empirical Comparison of Supervised Learning Algorithms.	Caruana & Niculescu-Mizil	comparison between ten supervised learning methods.	Using a variety of performance metrics/ Calibration Methods	-----	There is significant variability across the problems and metrics.	----- --	-----
19	2006	Link Prediction using Supervised Learning	Al Hasan et al	To study link prediction as a supervised learning task.	link prediction	Link -based Ranking	Data noisy, attribute values could be unknown.	Social networks	To consider time domains within number of data sets to understand link prediction.

Appendix D

Summary of literature on link mining and link mining techniques

20	2006	A latent Dirichlet model for unsupervised entity resolution	Indrajit & getoor	A probabilistic model for collective entity resolution for relational domains.	Latent Dirichlet Allocation (LDA)	Object related tasks. Object identification(entity resolution)	it does not make pair-wise decisions and introduces a group variable to capture relationships between entities	Bibliographic datasets.	Extending the model to resolve multiple entity classes.
21	2006	connecting SRL and Multi-Agent System (MAS)	Desjardins & Gaston	Relationship between (SRL)+(MAS)	Categorization of LM task: link-based classification, link-based ranking	Link prediction	MAS contribute to SRL.	focused on distributed methods that may be useful for scaling up SRL to large, complex networks.	Distributed methods for scaling up SRL to large and complex networks
22	2007	temporality in link prediction: understanding social complexity	potgiel. et al	to found that existing graph generation models are unrealistic	Dynamic Bayesian Networks (DBN).	generation models for graph	Temporal metrics are extremely contribution to link prediction.	Social network	Predicting relationships of time graphs density.

Appendix D

Summary of literature on link mining and link mining techniques

23	2007	Combining collective classification and link prediction.	Mustafa et al	General approach for combining object classification and link prediction.	Iterative Collective Classification and Link Prediction (ICCLP)	Collective Classification/link prediction	Attribute noise, link noise, link density	graphs	Exploring the other variations for combining collective object classification and link prediction.
24	2007	Predicting Structured Data	Bakr et al	To reduce the exponentially growing complexity with the label length.	Conditional Random Fields (CRF)	Generate models for graphs	How to provide much more accurate predict the labels of new samples.	the multi-label classification problem	the connection between the Conditional Graphical Models and the probabilistic approaches for solving the multi-label problem.
25	2007	Generating Social Network Features for Link-based Classification	Karamon et al	to bridge the gap between the aggregated features from the network data and traditional indices used in social network analysis.	classification	link-based classification,	The ratio of values, which has not been well investigated in sociology studies	social network	To encourage the application of KDD techniques to social Sciences, and vice versa.

Appendix D

Summary of literature on link mining and link mining techniques

26	2007	Collective Entity Resolution in Relational Data	BHATTACHARYA & GETOOR	propose a novel relational clustering algorithm that uses both attribute and relational information for determining the underlying domain entities, and implementation	clustering algorithm	Entity resolution / (object identification)	the gains diminish as relational patterns become less informative	multiple real-world databases	To study the algorithms on different types of relational data including consumer data, social network data, and biological data.
27	2008	Learning directed probabilistic logical model from relational data.	Daan Fierens	Directed probabilistic logical models	First-Order logic / Probabilistic logic models	Modeling logical vs statistical dependences.	non-recursive	Relational data	To learn useful recursive dependencies.
28	2009	Learning link – based classifiers from ontology-extended textual data.	Caragea et al	Addressing the problem of learning classifiers from structured relational data.	Learning link-based naïve Bayes classifiers on a text classification task/. Statistical methods "shrinkage".	Link- based classification	How semantically disparate relational data sources.	Relational data	Exploring the effect of using ontology's and mapping incompleteness and errors.

Appendix D

Summary of literature on link mining and link mining techniques

29	2009	Entity Linking through Neighborhood Comparison and RandomWalks	Liu	Explore two kinds of methods for the entity linking task. To compares the similarity between entities by their common neighbours, and second is asked on random walk models.	Neighbourhood Comparison & random walk	Entity resolution / (object identification)	Data size	links on Wikipedia page	The experiments on large scale data are left for further investigation.
30	2009	RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis	Sun et al	Address the problem of generating clusters for a specified type of objects, as well as ranking information for all types of objects based on these clusters in a multi-typed	K-clustering /graph clustering methods	Link based-rank/ Object clustering (group detection)	Research is needed to further consolidate this interesting framework and explore its broad.	(Heterogeneous) information network.	How to add citation information and text information to the bibliographic data? The empirical rules and its associated weight computation formulas proposed in this study may not be directly. the quality of ranking function is important to the accuracy of clustering, as it can capture the distinct feature for clusters.
31	2010	Probabilistic Similarity Logic	Bröocheler et al	Introduces probabilistic similarity logic (PSL), a general-purpose framework for joint reasoning about similarity in relational domains & incorporates probabilistic reasoning about similarities and relational structure in a principled way.	statistical relation learning(SRL)/ probabilistic similarity logic (PSL),	Link- based clustering(Group detection)	-----	Multi-relational data.	Studying different distance from satisfaction functions, such as L2 distance and applying PSL to other domains.

Appendix D

Summary of literature on link mining and link mining techniques

32	2010	A Theoretical Approach to L. Mining for personalization	Srinivas et al.	To a general Web search engine, and collect a number of the highest-scoring URLs.	-----	-----	The problem of query classification is extremely difficult owing to the brevity of queries	Data mining	D.M challenges in l. mining such as identify of the, Link discovery, common relational patterns.
33	2010	Entity Linking Leveraging Automatically Generated Annotation	Zhang et al	To use additional information sources from Wikipedia to find more name variations for entity linking task	A binary classifier based on Support Vector Machine (SVM)	Link- based ranking	It is difficult for the ranking approach to detect a new entity that is not present in KB, and it is also difficult to combine different features	Health care company	To improvements accuracy on KBP
34	2011	Meta Similarity Noise-free Clusters Using Dynamic Minimum Spanning Tree with Self-Detection of Best Number of Clusters	Karthikeya n&peter	A Minimum Spanning Tree based clustering algorithm for noise-free or pure clusters.	cluster /the DGEMSTNFM C algorithm	Met –data discovery	-----	database	To explore and test clustering algorithm in various domains. Find Best number of Meta similarity noise-free clusters to solving different clustering problems.

Appendix D

Summary of literature on link mining and link mining techniques

35	2011	Supervised Random Walks: Predicting and Recommending Links in Social Networks	Backstrom & Leskovec	To combine the information from network structure with rich node and edge attribute data remains largely open.	Supervised RandomWalks.	Based on feature extraction.	-----	social networks	To apply to many other problems that require learning to rank nodes in a graph, like recommendations, anomaly detection, missing link,
----	------	---	----------------------	--	-------------------------	------------------------------	-------	-----------------	--

Anomalies in Link Mining Based on Mutual Information



Research Student
Zakia I El Agure

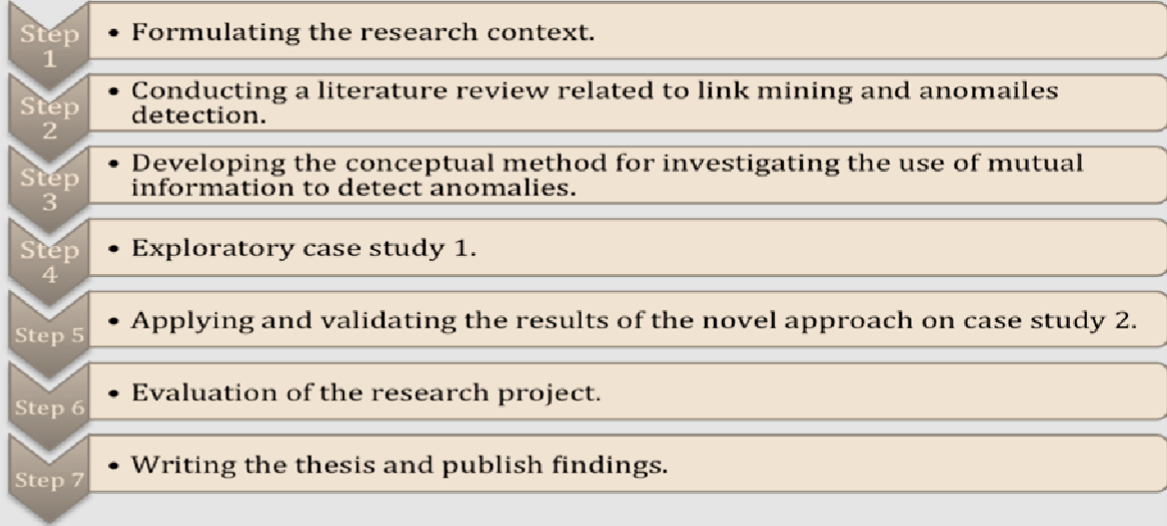


Domain

Link mining which is a new emerging research area, considers datasets as a linked collection of interrelated objects; it focuses on discovering explicit links between objects. Anomalies detection which is the focus of this research is concerned with the problem of finding anomalous patterns in datasets which can include outliers, exceptions, aberrations, surprises, or peculiarities (Chandola *et al.*, 2009).

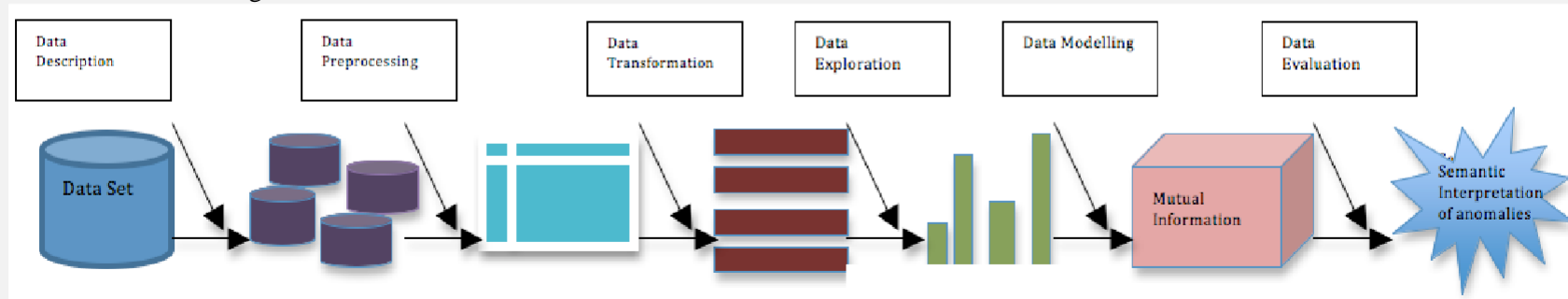
Aim

The aim of the research is to develop a novel approach to provide a semantic interpretation of anomalies based on mutual information in link mining.



Link Mining Methodology

As CRISP–DM methodology is well developed and applied in knowledge discovery, this research has adapted it to the emerging field of link mining. While data mining addresses the discovery of patterns in data entities, link mining is interested in finding patterns in objects by exploiting and modelling the link among the objects. The approach to link mining is still an ad-hoc approach. The proposed adopted CRISP-DM methodology can help provide a structured approach to link mining. This consists of six stages:



Original Contributions

Major Contributions:

- 1- Use of anomalies in link mining.
- 2- Using MI to provide semantic interpretation of anomalies.
- 3- Case Study 2 demonstrates that MI can be used to validate the clustering and visualisation.

Minor Contributions:

- 1-Modified CRISP to support link mining.
- 2-Applied CRISP to support the use of MI for anomaly interpretation.

Experimental Study

1. The first case study is used as proof of concept to examine the validity of the proposed approach. The mutual information is applied to case 1 to understand/explain anomalies approach.
2. The second case study to demonstrate how mutual information can help explore and interpret the anomalies detection in link mining. The development of novel techniques for link mining is the key challenge for this technique to make use of the same approach to a different real world data set, to a different form of data representation based on graphs using different clustering approach (hierarchical cluster) as this validates the approach through visualisation.

Appendix F: Mind Map

